Titre: Title:	Science des données et politique : quatre essais pour comprendre les processus démocratiques
Auteur: Author:	William Sanger
Date:	2019
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Sanger, W. (2019). Science des données et politique : quatre essais pour comprendre les processus démocratiques [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/3872/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/3872/
Directeurs de recherche: Advisors:	Nathalie de Marcellis-Warin, & Thierry Warin
Programme: Program:	Doctorat en génie industriel

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Science des données et politique : quatre essais pour comprendre les processus démocratiques

WILLIAM SANGER
Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie industriel

Avril 2019

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

Science des données et politique : quatre essais pour comprendre les processus démocratiques

présentée par **William SANGER**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Mario BOURGAULT, président
Nathalie DE MARCELLIS-WARIN, membre et directrice de recherche
Thierry WARIN, membre et codirecteur de recherche
Marcelin JOANIS, membre
Florence MILLERAND, membre externe

DÉDICACE

Les mots ne sauraient être suffisants pour remercier Gabrielle Crétot-Richert qui partage ma vie depuis les bancs universitaires. Cette complicité s'est transformée en odyssée passionnante, où les découvertes et les voyages font place à un futur que l'on construit ensemble jour après jour.

Le 2 septembre 2017, j'ai eu l'honneur d'être accueilli au sein de la famille mécano. Votre bienveillance ont façonné la personne que je suis. À Hélène Boisvert pour son amour, Serge Occhietti pour sa confiance, Raphaelle Occhietti et Stefano Menegat pour leur amitié.

Finalement, à ma grand-mère, Maguy Sanger, qui pétille de vitalité et de tendresse. Ma tante et exploratrice, Silvia Sugihara, qui m'appuie sans question. Mes frères, Alexandre et Edward, toujours présents. Et mon père, Gérard Sanger, qui aurait lu avec fierté et malice cette thèse en repensant aux disquettes de jeu de mon enfance.

REMERCIEMENTS

L'aventure doctorale est un voyage aux multiples étapes. C'est un parcours jonché de rencontres, de découvertes intellectuelles et de remises en question. À une époque où l'instantanéité prime, un doctorat est une occasion unique d'observer le monde autour de soi, et de marquer le temps qui passe. Je tiens donc à prendre ces quelques lignes pour remercier les personnes qui m'ont soutenu au cours de ces années et qui parsèment mon itinéraire.

D'abord mes directeurs de thèse, Nathalie de Marcellis-Warin et Thierry Warin. Vous m'avez permis de goûter aux joies de la recherche académique et de me dépasser au travers de nombreux projets ambitieux. J'ai notamment pu développer mes compétences sur votre plateforme de recherche, Nüance-R, afin de m'exprimer en tant que scientifique de données.

Le CIRANO et le Mondo Lab/Social Data Science Lab, pour avoir accueilli mes recherches depuis 2011. Je tiens à remercier particulièrement Ghislain Camirand et Jérôme Blanc pour leur aide sans faille. Aussi HEC Montréal, pour m'avoir fourni un espace de travail.

Polytechnique Montréal, et les cohortes d'étudiants que j'ai eu le privilège d'encadrer. J'ai aussi eu le plaisir de suivre des cours passionnants à HEC Montréal (Thierry Warin), à l'Université de Montréal (André Blais) et à l'UQAM (Florence Millerand) pendant le doctorat.

Gaëtan Madiès, unis par les campagnes électorales, les voyages et les mariages ; Marco Lugo, pour avoir hacker la démocratie ; Frédéric Séguin, compagnon depuis Prague et à travers les Éditions Sans G ; Marine Hadengue, Yoann Guntzburger, Julie Charron-Latour et Noémie Cabau, pour partager les épreuves émotives du doctorat ; Sandrine Reny, qui n'a jamais été aussi proche malgré les frontières ; Iva Zajacova et Ladislav Vlček, pour leur amitié. Matthieu Bister et sa petite famille. Serge Occhietti, pour avoir relu à maintes reprises cette thèse.

Au sein du CIRANO a été tissé le Social Data Science Lab, où j'ai croisé de brillants collègues : Antoine Troadec, Bertrand Nembot, Farnaz Farnia, Romain Le Duc, Charlotte Vorreuther, Marine Leroi, Johanne Basile, Thibault Sénégas, Julie Wojcicki et Christophe Mondin. La camaraderie, leur aide et les défis intellectuels étaient toujours au rendez-vous.

Cette recherche n'aurait pu être menée sans la générosité du Conseil de Recherche en Sciences Humaines (CRSH), son concours de vulgarisation scientifique, la bourse J.A. Bombardier de Polytechnique Montréal et le prix de la Relève Universitaire de Finance Montréal.

Je tiens à remercier trois professeurs qui sans le savoir ont influencé ma destinée : François Pereira (École Jean Achard), Fabien Hulot (Collège Stanislas) et Pierre Savard (Polytechnique Montréal). Merci d'avoir cru en moi, cette thèse est le résultat de votre pari.

RÉSUMÉ

Les comportements politiques des citoyens, des partis politiques et des institutions démocratiques ont évolué depuis la naissance d'Internet. Aucune élection ne se passe désormais sans l'utilisation de données générées par les individus, que ce soit avec des sondages ou avec des interactions sur les médias sociaux. En parallèle, de nouvelles méthodologies quantitatives permettent d'interpréter ces nouvelles données.

Cette thèse doctorale se concentre sur la question de recherche suivante : comment les données massives et la science des données peuvent être utilisées pour comprendre les processus démocratiques à l'ère d'Internet ?

Après une revue algorithmique de la littérature académique concernant les sciences politiques et les nouvelles données, puis le développement de la littérature associée aux sciences politiques et aux médias sociaux, quatre pistes de recherche sont explorées, permettant de répondre à la question de recherche générale. Chacune est associée à un article de recherche constituant le corps de la thèse. Les données récoltées à travers cette thèse doctorale sont principalement issues de Twitter (articles 1 à 3). L'article 4 utilise l'ensemble des manifestes politiques européens entre 2000 et 2018. Concernant la méthodologie, cette thèse doctorale repose sur la science des données (acquisition de données massives à partir de réseaux sociaux, économétrie, visualisations de données, traitement automatique du langage naturel).

Le premier article se rapporte aux élections québécoises de 2014, et décrit la campagne électorale perdue par le Parti Québécois malgré le fait que le parti ait été au pouvoir au moment du déclenchement de l'élection. L'utilisation de modèles économétriques a permis d'associer préférentiellement les thématiques de campagne aux quatre chef·fe·s de partis politiques à partir de 672 497 tweets.

Le deuxième article de recherche prend pour terrain d'observation l'élection fédérale canadienne de 2015. Les techniques d'analyse textuelle ont permis de traiter près de 3,5 millions de tweets et de révéler les dynamiques de campagnes menant à la victoire du Parti Libéral du Canada.

Le troisième article de la thèse met en oeuvre plusieurs modèles économétriques pour étudier plus de deux millions de messages publiés sur Twitter au cours de la campagne électorale nigériane de 2015. Ces techniques mettent en perspective l'utilisation des données issues des médias sociaux comme source supplémentaire d'informations pour consolider la portée des sondages traditionnels.

Finalement, le quatrième article de la thèse se concentre sur les différentes élections européennes ayant eu lieu entre 2000 et 2018. À partir d'une base de données de 12 millions de mots, la création de nouveaux indicateurs mesurant la similarité entre les partis politiques permet d'appréhender la notion de populisme à travers les pays européens.

Les contributions de la thèse sont de trois natures. (1) Méthodologiquement, cette thèse met en oeuvre de nombreuses techniques en science des données. Cela va de la collecte de données inédites issues des médias sociaux, à la création de nouveaux indicateurs de suivis électoraux, jusqu'à la comparaison de documents écrits en plusieurs langues ou à l'attribution de thématiques de campagne aux différents candidats grâce à des modèles économétriques ou des techniques d'apprentissage semi-supervisé. Ces méthodologies permettent de comprendre le déroulement d'une élection moderne alors que sont générées en temps réel les données des individus et des organisations. (2) Les contributions sont aussi de nature théorique, avec la caractérisation des partis de gouvernement par rapport aux partis extrêmes et l'étude du populisme. (3) Finalement, les contributions sont de nature thématique, avec la publication de recherches concernant les élections québécoises de 2014, canadiennes de 2015, nigérianes de 2015 et européennes entre 2000 et 2018.

ABSTRACT

Political behaviour of citizens, political parties and democratic institutions have evolved since the advent of the Internet. Nowadays, no election takes place without the use of real-time data provided by individuals. At the same time, new quantitative methodologies are being used to interpret these data.

This doctoral thesis focuses on the following research question: how unstructured data (big data) and data science can be used to understand democratic processes in the Internet age?

After an algorithmic review of the academic literature and an analysis of the literature associated to political science and social media, four research avenues are outlined to answer the general research question. Each one is associated to a research article as the corpus of the thesis. The data collected through this doctoral thesis are mainly from Twitter (articles 1 to 3). Article 4 takes into account all European political manifestos between 2000 and 2018. With regard to the methodology, this doctoral thesis is based on data science (data acquisition from social networks, econometrics, data visualization, algorithmic analysis of textual content).

The first article refers to the 2014 Quebec election, and describes the election lost by the Parti Québécois despite having been the incumbent party. The use of econometric models made it possible to associate campaign topics with the four leaders of political parties from 672,497 tweets.

The second research article takes as its field of observation the 2015 Canadian federal election. With textual analysis techniques, nearly 3.5 million tweets have been processed in order to understand the victory of the Liberal Party of Canada.

The third article of the thesis uses several econometric models to study more than two million messages published during the 2015 Nigerian election campaign. These techniques put in perspective the use of social media data as an additional source of information to consolidate traditional surveys.

Finally, the fourth article of the thesis focuses on the various European elections that took place between 2000 and 2018. Using a 12 million word database, the creation of new indicators measuring similarity between political parties makes it possible to understand the notion of populism across European countries.

The contributions of the thesis are of three types. (1) Methodologically, this thesis uses several techniques in data science. This ranges from collecting new data from social media, to

creating indicators monitoring elections, to comparing documents in different languages or assigning campaign themes to candidates using econometric models or semi-supervised machine learning techniques. These methodologies make it possible to understand the conduct of a modern election when data from individuals and organizations are generated in real time. (2) The contributions are also of a theoretical nature, with the characterization of governing parties versus populist parties. (3) Finally, the contributions are of a thematic nature, with the publication of research on the Quebec elections of 2014, the Canadian elections of 2015, the Nigerian elections of 2015 and the European elections between 2000 and 2018.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

APC	All Progressive Congress (Nigeria)
API	Application Programming Interface
CAQ	Coalition Avenir Québec
CEO	Chief Executive Officer
FERA	Federal Emergency Relief Administration
FN	Front National
FTA	Free Trade Agreement
IDE	Integrated Development Environment
LDA	Latent Dirichlet Allocation (allocation de Dirichlet latente)
LREM	La République En Marche (France)
MAE	Mean Absolute Error
MoDem	Mouvement Démocrate
NPD	Nouveau Parti Démocratique (Canada)
ONG	Organisation non gouvernementale
PCC	Parti Conservateur du Canada
PDP	People's Democratic Party (Nigeria)
PLC	Parti Libéral du Canada
PLQ	Parti Libéral du Québec
PQ	Parti Québécois
QLP	Québec Liberal Party
QS	Québec Solidaire
RT	Retweet
TF-IDF	Term Frequency-Inverse Document Frequency
UDI	Union des Démocrates Indépendants
UKIP	UK Independence Party
UMP	Union des Mouvements Populaires (France)

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La chute du Mur de Berlin coïncide à quelques années près au développement du réseau Internet. À l'observateur politique de l'époque, notre environnement ne peut qu'être plus différent. Trois questions pourraient alors être posées vingt ans après le début du XXI^{e} siècle : comment évolue le paysage démocratique, comment se déroule une élection à l'ère d'Internet, et comment se structurent les interactions entre citoyens et représentants politiques ?

Déjà cinq siècles avant l'introduction des premiers téléphones intelligents, Nicolas Machiavel étudiait la relation entre peuple et Prince, et mentionnait que "dès qu'on est élevé sur le trône par la faveur du peuple, il est absolument nécessaire de s'en faire aimer, ce qui est extrêmement aisé, car il n'exige rien que de n'être pas opprimé" [Machiavel, 1532].

Époque plus complexe qu'est la nôtre en comparaison. Tandis que les tweets et les *posts* tissent des toiles d'informations entre individus, les politicien·ne·s sont de plus en plus scruté·e·s en continu. Qu'hier ce fusse à travers les journaux ou la radio, puis la télévision, et aujourd'hui sur Internet, le flux d'informations politiques ne semble pas tarir. À l'heure actuelle, jamais citoyen·ne n'a eu autant accès à un réseau d'informations.

Mais remontons encore un peu plus loin. 2 500 ans séparent le klérotèrion athénien de la succession immuable de la blockchain. Le premier servait d'outil de désignation des citoyens pour la participation aux assemblées publiques. De manière transparente, aléatoire, chaque citoyen prenait part à tour de rôle de manière aléatoire aux discussions de la cité athénienne. Transparente. Tout comme la chaîne des blocs qui s'agrège en temps réel, offrant un tracé indélébile et incorruptible relatant les interactions entre utilisateurs. Cette technologie pourra servir de colonne vertébrale aux prochaines applications démocratiques.

Depuis une dizaine d'années surgissent telle une lame de fond plusieurs changements structurels importants dans nos sociétés. D'un côté, la démocratisation de la technologie et l'accès à Internet modifient la manière dont les individus interagissent avec l'information. Jamais autant de données n'ont été disponibles pour comprendre un phénomène, observer les liens entre personnes, décrypter l'environnement dans lequel nous vivons. Toutefois, un paradoxe émerge puisque les perceptions des individus semblent isoler progressivement des pans de la société au complet. Ainsi, le mur de la caverne de Platon semble de plus en plus éloigné de l'ouverture du monde réel, à une époque où les campagnes électorales restent criblées par la propagande algorithmique et où les citoyens se regroupent en poches homophiles aux affinités similaires.

Est-ce un phénomène nouveau ? Livres, journaux, radio, télévision, Internet, téléphones intelligents et médias sociaux témoignent de la propagation du savoir et de la culture. Le dernier de la liste n'est-il qu'un simple prolongement de phénomènes déjà existants ou une redéfinition de dynamiques complexes que les sociétés du XXI^{e} siècle devront prendre en compte ?

Populisme et propagande ne sont toutefois pas des phénomènes nouveaux. Alors que les discours politiques étaient auparavant débattus ou appuyés par un pouvoir médiatique à travers la presse, la télévision ou la radio, on assiste à une migration des discours politiques vers des plateformes non cloisonnées telles que YouTube, Facebook ou Twitter.

À l'instar de l'ouverture des frontières territoriales, les conséquences des élections deviennent de plus en plus globales. Ainsi de fausses informations propagées lors d'un scrutin électoral se répercuteront sur les partenaires commerciaux de ce pays, sur les entreprises faisant partie des différentes chaînes de valeurs, sur les travailleur·euse·s situé·e·s à des milliers de kilomètres des urnes de vote. À la promesse du village global que fut utopiquement Internet, on remarque un déplacement vers des chambres d'écho devenant de plus en plus imperméables ; toutefois, les impacts traversent les frontières, les données n'ayant plus de nationalité à l'ère de la post-vérité.

Les inégalités et les nationalismes ont augmenté depuis la crise financière de 2008. Durant la même période, les individus se sont organisés physiquement et virtuellement grâce à Internet et plus particulièrement grâce aux médias sociaux. Le(s) printemps arabe(s), les mouvements du type Occupy Wall Street, les manifestations de Hong-Kong (révolution des parapluies), le rejet des institutions européennes et plus récemment la crise des migrants en Europe ne sont que quelques exemples frappants de l'utilisation de ces nouvelles données.

En parallèle, les institutions démocratiques ont dû s'adapter à de tels changements. Les élections évoluent et se transposent dans un état continu où les électeur·ice·s sont sollicité·e·s en continu. L'élection de Barack Obama en 2008 a pavé la voie à de nouvelles stratégies électorales que tous les partis politiques modernes suivent désormais. Les méthodes traditionnelles de sondage de l'opinion ne montrent toutefois pas le même niveau de certitude face à l'évolution des campagnes électorales : lors de l'élection législative britannique de 2015, aucun sondage n'a pu prévoir la victoire du parti de David Cameron. En effet, pendant les 48 mois précédant le scrutin, le chef du Parti Conservateur était constamment en deuxième position face au Parti Travailliste d'Ed Miliband. L'utilisation de données générées en temps réel par l'ensemble des individus permettrait d'éviter de telles situations, notamment en comprenant avec plus de finesse les avis des citoyens, ou du moins réduire les asymétries d'information.

Finalement, tandis que les citoyen·ne·s ont accès à une information sans commune mesure, on

assiste aussi à certain effritement démocratique. L'offre politique populiste fleurit à travers le monde, avec une remise en cause des institutions. De plus, la participation électorale est en baisse au cours des dernières décennies. À titre d'exemple, le taux de participation aux élections européennes était légèrement au-dessus de 60% en 1978, et baissa progressivement pour atteindre moins de 45% en 2014.

Les plateformes numériques occupent une place prépondérante dans le débat démocratique, où partis politiques, journalistes et citoyen·ne·s échangent sur Facebook, Twitter ou YouTube. Les tactiques d'utilisation de ces plateformes ont été maîtrisées rapidement par les partis politiques et les gouvernements.

Le terme de post-vérité a émergé durant les cinq dernières années, et le contenu à teneur politique disponible en ligne semble impossible à filtrer. Google précise dans une entrevue au Guardian en décembre 2016 que l'information obtenue en effectuant une requête est un reflet du contenu du web, que ce contenu soit véridique ou non. À la post-vérité se sont ajoutées les "fake news", ou fausses nouvelles. Lors du vote du Brexit en 2015 et lors de l'élection de Donald Trump à la présidence des États-Unis en 2016, la compagnie d'analyse de données massives Cambridge Analytica a mis des méthodologies de traitement de données au service des partis politiques respectifs, afin de promouvoir un message calibré selon leurs revendications. L'observateur ayant assisté à la chute du Mur de Berlin n'aurait pu anticiper autant d'effets de la venue d'une technologie telle qu'Internet sur la démocratie, avec ses impacts positifs et négatifs. Nous vivons dans une nouvelle ère caractérisée par de nouveaux outils et présentant de nouvelles dynamiques.

Et c'est là l'objectif et la question de recherche de cette thèse doctorale : comprendre **comment les données massives et la science des données peuvent être utilisées pour interpréter les processus démocratiques à l'ère d'Internet ?**

1.1 Plan de la thèse

Afin de répondre à cette question de recherche, la thèse doctorale est divisée en plusieurs parties. Le deuxième chapitre de la thèse est dédié à la revue critique de la littérature académique. Celle-ci est regroupée en cinq sections, dont la première est une revue de littérature algorithmique considérant l'ensemble de la littérature académique disponible. Viennent par la suite l'analyse des forces et comportements politiques, des médias de masse, de la représentation des populations et des données massives en politique.

Le troisième chapitre de la thèse fait écho aux nouvelles questions de recherche soulevées par la revue de la littérature. Il présente la question de recherche centrale de cette thèse, les

hypothèses abordées par les quatre articles de recherche, ainsi que la méthodologie utilisée concernant les données récoltées, les modèles économétriques calibrés et les techniques de science des données.

Le corps de la thèse est divisé en quatre chapitres, chacun représentant un article de recherche. Ainsi, le premier article de la thèse est consigné au chapitre quatre. Intitulé *The Public's Perception of Political Parties During the 2014 Quebec Election*, il présente l'analyse des messages issus de Twitter durant l'élection québécoise de 2014.

Le second article de la thèse, intitulé *The 2015 Canadian Election on Twitter : A Tidy Algorithmic Analysis* présente l'analyse des données récoltées sur Twitter durant l'élection canadienne de 2015 et fera l'objet du cinquième chapitre de la thèse.

Le chapitre six accueille le troisième article de la thèse *Nigeria's 2015 Presidential Election : A Spatial and Econometric Perspective based on a Framing Strategy* et fait état de l'élection de 2015 ayant eu lieu au Nigeria.

Le dernier article de la thèse se trouve au chapitre sept et présente l'analyse des manifestes de partis politiques européens depuis 2000. Cet article s'intitule *Text-as-Data Analysis of Populist Parties versus Government Parties : to Blend or not to Blend ?*

Au chapitre huit sera présentée une discussion générale, mettant en perspective les différents résultats obtenus à travers les articles de recherche par rapport à la littérature académique exposée.

Finalement, la thèse se clôture au chapitre neuf avec le détail des apports théoriques, méthodologiques et thématiques, mais aussi les limites et les recommandations de ce travail.

CHAPITRE 2 REVUE DE LA LITTÉRATURE

Cette thèse s'ancre dans une littérature académique prolifique et en mutation. L'utilisation de nouvelles techniques d'analyse permet de mesurer les comportements des individus. Les données, de plus en plus importantes en volume et de plus en plus précises par leur granularité, offrent des champs de travail complexes et originaux. Cinq axes de réflexion seront développés dans cette section, dressant le cadre analytique utilisé par cette thèse doctorale.

Traditionnellemnt, une revue littérature systématique permet de paver la voie à une recherche éclairée sur les tendances d'un champ académique. À partir d'un nombre important d'articles, et suite à des critères de plus en plus précis, le panel d'articles à analyser reste ambitieux mais réalisable et accessible pour le chercheur. Le prisme d'analyse choisi ne prend donc pas en compte l'entièreté de la littérature existante, à cause de la contrainte temporelle d'une lecture exhaustive des articles scientifiques. Bien qu'étant un processus reproductible, le nombre d'articles, et donc le nombre de références pouvant être utilisées, restent limités.

Dans un premier temps, nous présenterons une revue de littérature "algorithmique" à partir de l'analyse de 703 425 articles scientifiques. Le but d'une revue de littérature "algorithmique" est de considérer l'ensemble de la littérature académique disponible à l'aide de techniques bibliométriques et d'algorithmes. Les techniques de fouilles de données et de visualisation de données seront utilisées afin de représenter la diversité d'un champ de recherche à la production prolifique. Après avoir aggréger les mots clefs concernant chacun de ces articles scientifiques, plusieures thématiques seront alors identifiées et isolées, ce qui consituera les axes d'analyse de notre revue de littérature. Cette thèse doctorale se positionne ma thèse au sein de quatre types de littératures.

Ainsi, une synthèse de travaux reliés aux comportements politiques et aux forces en action lors des élections constituera la seconde partie de la revue de littérature. Cette section s'intéressera aux relations de pouvoir entre partis politiques, individus et institutions. Les médias de masse et leur influence en politique occuperont une troisième section. Viendra aussi la question de représentativité des populations à travers les sondages et dans les études utilisant des données massives. La cinquième partie traitera des dynamiques électorales en lien avec l'utilisation de nouvelles données lors des élections depuis l'avènement des médias sociaux. Finalement, nous dresserons les nouvelles questions de recherche issues de cette littérature.

2.1 Revue de littérature algorithmique

Le but de cette section est de comprendre la diversité des champs disciplinaires dans lesquels la thèse doctorale puisera sa littérature académique. Les données proviennent de la base de données académique Web of Science, gérée par la compagnie Clarivate Analytics (anciennement Thomson Reuters). Bien que non-exhaustive, cette base de données fait appel à six ressources principales couvrant une large portion de la production académique mondiale :

- Web of Science Core Collection : sciences, sciences sociales, arts et sciences humaines (depuis 1989)
- MEDLINE : biomédecine, santé publique, soins cliniques, sciences animales et végétales (depuis 1950)
- KCI-Korean Journal Database : accès aux journaux multidisciplinaires en Corée (depuis 1980)
- SciELO Citation Index : sciences, sciences sociales, arts et sciences humaines en Amérique Latine (depuis 1997)
- Russian Science Citation Index : science, technologie, médecine, éducation en Russie (depuis 2005)
- Derwent Innovations Index : brevets déposés à travers 50 institutions nationales ou internationales (depuis 1963)

De part la nature mixte de l'objet d'étude de cette thèse doctorale, trois sections seront abordées avec cette méthodologie algorithmique : (1) les sciences politiques, (2) les médias sociaux et les données massives et (3) l'intersection de ces deux champs de recherche.

Le but de cette section est de prendre en compte l'ensemble des articles scientifiques et d'identifier les champs de recherche associés, la provenance des publications académiques et finalement la mise en valeur des thématiques principales de la littérature. À nouveau, cette méthodologie a pour bénéfice de considérer un nombre d'articles scientifiques impossible à lire dans une carrière scientifique, d'où l'importance de l'utilisation de techniques algorithmiques.

2.1.1 Sciences politiques

Au total, 703 425 articles scientifiques sont recensés dans les bases de données de Web of Science concernant les sciences politiques et les élections. Les termes de recherche utilisés dans la section thématique sont "politic*" et "election", soit les expressions régulières faisant référence aux sciences politiques et aux élections en général. Cette recherche et les subséquentes ont été réalisées en date de juillet 2018.

De ce total de références, 572 888 proviennent de la collection principale de Web of Science,

79 441 de MEDLINE, 52 264 de KCI Korean Journal Database, 46 273 de SciELO Citation Index, 10 269 du Russian Science Citation Index et 6 930 de Derwent Innovations Index. De plus, 497 113 sont associées aux sciences sociales, 262 742 aux technologies et 193 348 aux sciences humaines. La figure 2.1 présente l'évolution du nombre de publications entre 1989 et 2017 (2018 n'étant pas encore une année complète dans les bases de données de Web of Science à l'écriture de la thèse).

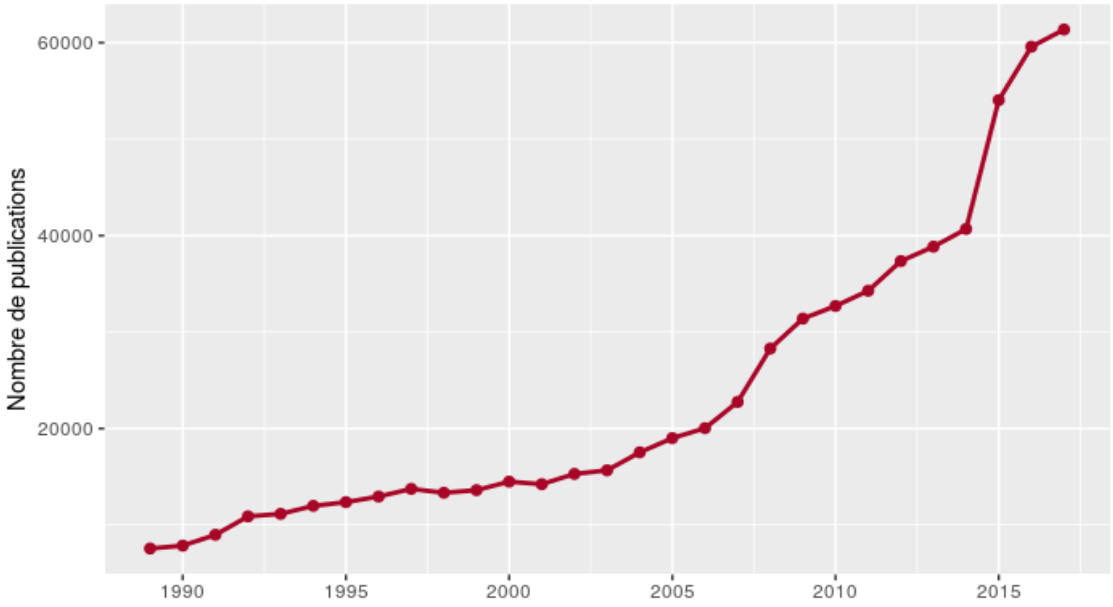

Figure 2.1 Nombre de publications recensées sur Web of Science 1989-2017

Parmi les 100 plus importantes zones de publications, les États-Unis arrivent en tête avec près de 30.6% des publications dans le domaine. Suivent le Royaume-Uni (14.45%), la Chine (4.4%), le Canada (4.34%) et l'Allemagne (4.03%). La cartographie suivante (figure 2.2) permet de visualiser ces zones de publication en fonction du nombre d'articles publiés recensés sur Web of Science.

Les dix domaines de recherche les plus représentés sont en ordre d'importance la gouvernance et le droit (187 416 publications sous la catégorie Law and Government), les affaires (122 391, Business), l'administration publique (109 004, Public Affairs), la sociologie (89 309), les problématiques sociales (81 913), l'histoire (81 450), les sciences sociales (76 099), les sciences informatiques (63 927), l'environnement (59 998) et la psychologie (53 549). La figure 2.3 présente les thématiques les plus représentées au sein de la base de données de Web of Science.

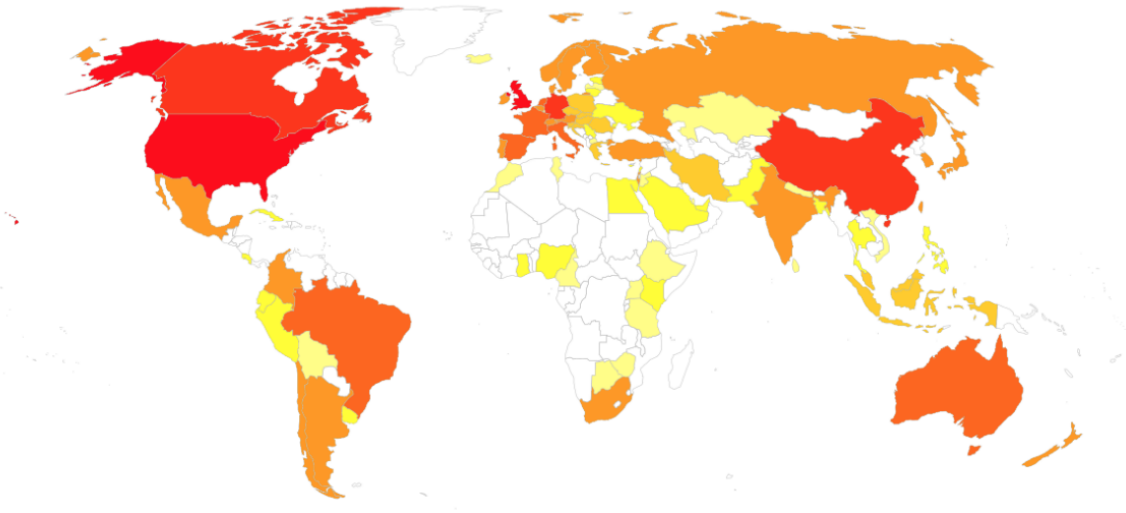

Figure 2.2 Cartographie des publications académiques mentionnant les sciences politiques et les élections

2.1.2 Médias sociaux, données massives

En prenant comme point de départ 1998, soit le premier article mentionnant le futur moteur de recherche qui est devenu par la suite Google [Brin et Page, 1998], 74 038 références académiques sont référencées dans Web of Science (année 2018 non complète). Pour cela, notre stratégie de recherche fut de sélectionner les articles ayant pour thématiques les éléments "social media" ou "big data". La progression des publications suit une trajectoire exponentielle de 2010 à 2016 (figure 2.4).

Si l'on observe les pays d'où proviennent ces publications, nous retrouvons à nouveau les États-Unis en première position avec près du quart des publications (23.5%). Viennent ensuite la Chine à 18.0%, la Grande-Bretagne (9.2%), l'Australie (4.2%) et le Canada (3.1%). L'ensemble des pays est représenté dans la cartographie de la figure 2.5.

Sans surprise, c'est dans le domaine des sciences de l'informatique et de l'ingénierie que se retrouvent la majorité des publications. Les télécommunications, la communication et les affaires clôturent les cinq domaines les plus représentés dans ce champ de recherche (figure 2.6).

2.1.3 Sciences politiques, médias sociaux et données massives

Dans cette dernière section, nous concentrons notre angle d'analyse sur l'intersection des deux champs de recherche précédents. Ainsi, les termes de recherche spécifiés sont doubles :

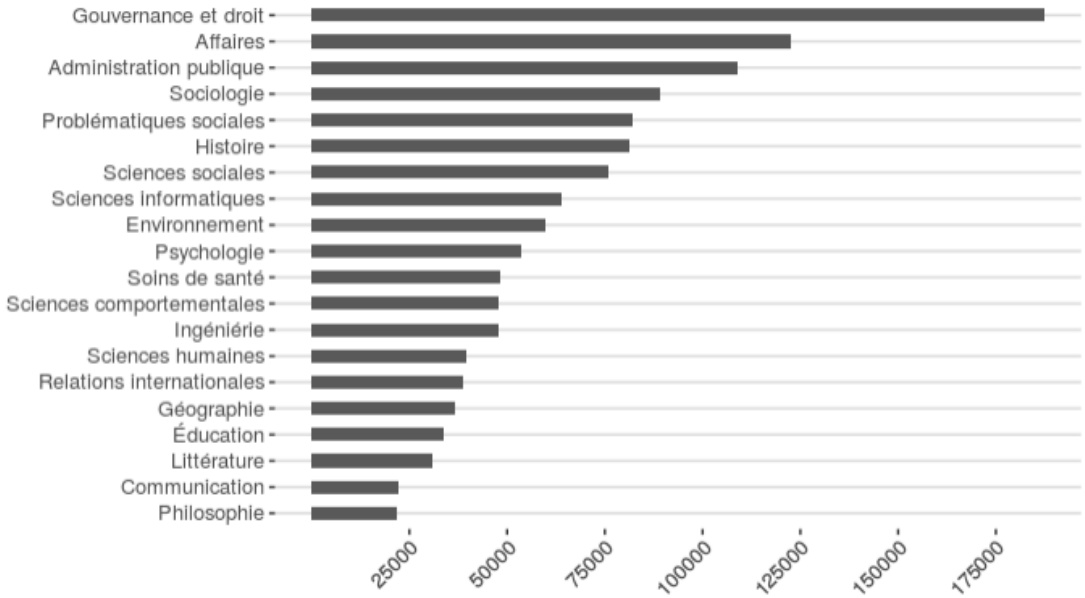

Figure 2.3 Disciplines académiques mentionnant les sciences politiques et les élections

d'une part les articles ayant pour thématique les données massives ("big data") ou les médias sociaux ("social media") et d'autre part les articles traitant de politique ou d'élections. Nous restreignons à nouveau les dates utilisées pour ne sélectionner que les articles publiés entre 1998 et 2018.

Au total, ce sont donc 5 044 articles qui sont disponibles à travers l'ensemble des bases de données de Clarivate Analytics. Un des avantages de l'utilisation de revue de littérature algorithmique est la capacité à faire émerger les nouveaux concepts présents dans une discipline académique, notamment par des techniques bibliométriques. Une telle méthodologie a été appliquée dans le domaine de l'intelligence artifielle où les interactions entre les différents sous-champs ont été observés [van Eck *et al.*, 2006].

Concernant donc le champ des sciences politiques, des médias sociaux et des données massives, aucun article avant 2004 n'est catalogué dans la base de données. Malgré le fait que Facebook voit le jour en 2004 et Twitter l'année suivante, il est intéressant de noter que le nombre de publications académiques connaît deux augmentations majeures, une première à partir de 2010 et la suivante depuis 2014 (figure 2.7).

Encore une fois, les États-Unis (25.6%), la Grande-Bretagne (14.1%), la Chine (8.2%), l'Australie (4.8%) et l'Allemagne (3.4%) figurent comme les pays d'où proviennent le plus de recherches à travers cette littérature. Il est à noter que cette fois-ci, bien que premier pays

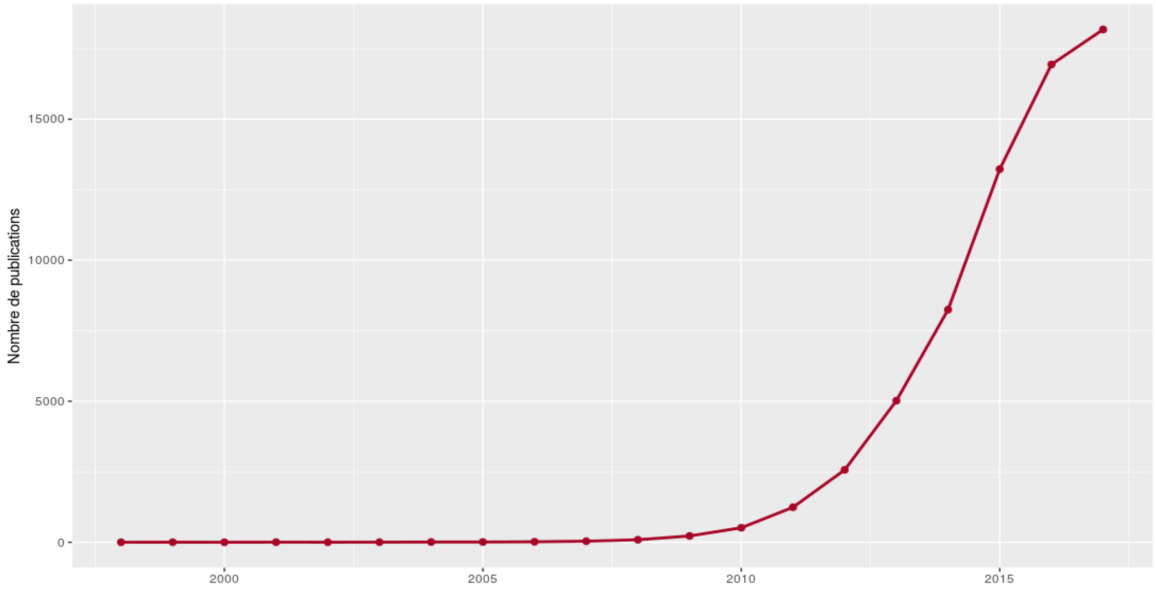

Figure 2.4 Nombre de publications recensées sur Web of Science sur le thème des médias sociaux et des données massives, 1998-2017

émetteur de publications, l'écart avec les autres pays de tête est toutefois moins important (figure 2.8).

Parmi l'ensemble des disciplines académiques, les domaines les plus représentés sont en premier lieu les sciences informatiques. Viennent ensuite la communication, les sciences comportementales, les sciences sociales et les problématiques sociales. La figure 2.9 regroupe les 20 thématiques les plus représentés dans ce champ de recherche.

En se concentrant seulement sur les articles catalogués par Web of Science, nous utiliserons une méthodologie issue des sciences de données pour considérer l'ensemble de la littérature académique à l'intersection entre les sciences politiques et les données massives. Principalement, nous répliquerons l'article de [Warin *et al.*, 2018] utilisant la méthodologie développée par [van Eck *et al.*, 2006]. Ces derniers ont analysé la discipline de l'intelligence artificielle, dévoilant les nouvelles tendances de recherche. Le logiciel VOSviewer est utilisé afin d'analyser les données bibliométriques à partir des mots clefs de chaque publication.

Pour obtenir une visualisation complète de la littérature utilisée, les paramètres d'utilisation du logiciel restent similaires à [Warin *et al.*, 2018]. Dans un premier temps les mots clefs de chaque article sont considérés. La figure 2.10 est obtenue suite à l'analyse de réseaux de l'ensemble des relations unissant ces mots clefs aux articles scientifiques. Afin de reproduire les résultats suivants, une méthode de normalisation par force d'association a été utilisée, avec

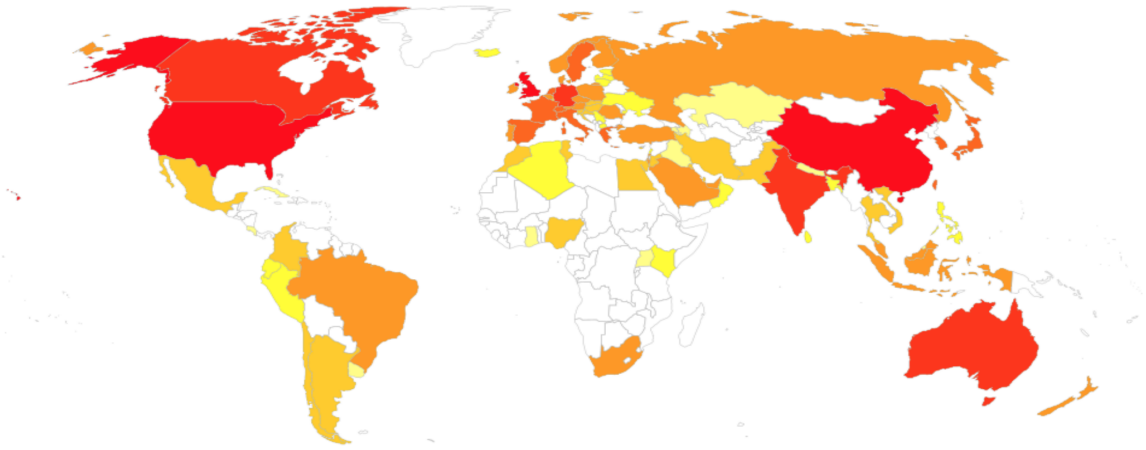

Figure 2.5 Cartographie des publications académiques mentionnant les données massives et les médias sociaux

des paramètres d'attraction et de rejet de 2 et 0, et finalement les plus petits regroupements de thématiques peuvent être fusionnés. Ces paramètres de visualisation permettent de favoriser la représentations de communautés.

Ainsi la figure 2.10 permet de représenter les mots clefs de la littérature à partir de ces 4 590 articles scientifiques. [1]

1. Ce nombre diffère des 5 044 articles précédemment mentionnés ; un des paramètres de visualisations de VOSviewer est la sélection du nombre de thématiques à représenter ; dans notre cas, les mots clefs ne concernant qu'un article de recherche ne sont pas considérés, d'où un nombre total d'article légèrement inférieur.

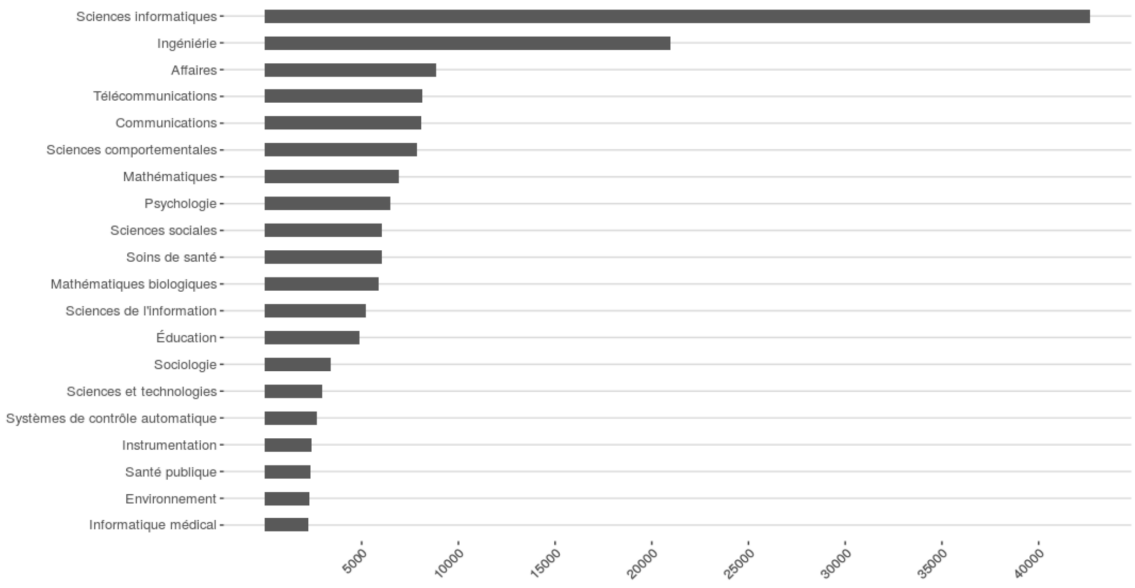

Figure 2.6 Disciplines académiques mentionnant les données massives et les médias sociaux

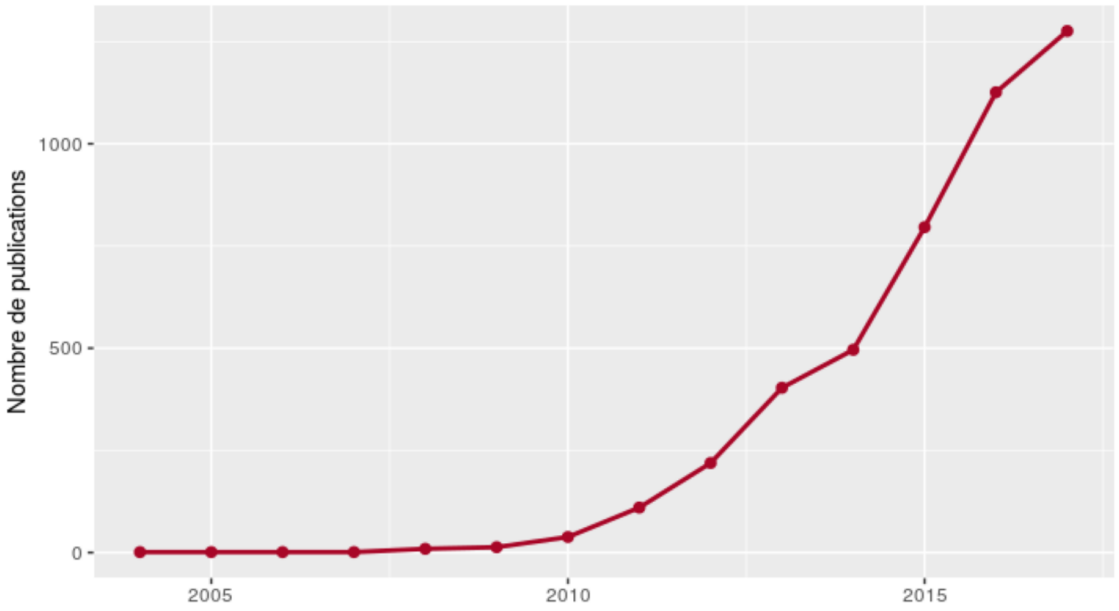

Figure 2.7 Nombre de publications recensées sur Web of Science 1989-2017 (sciences politiques, médias sociaux et données massives)

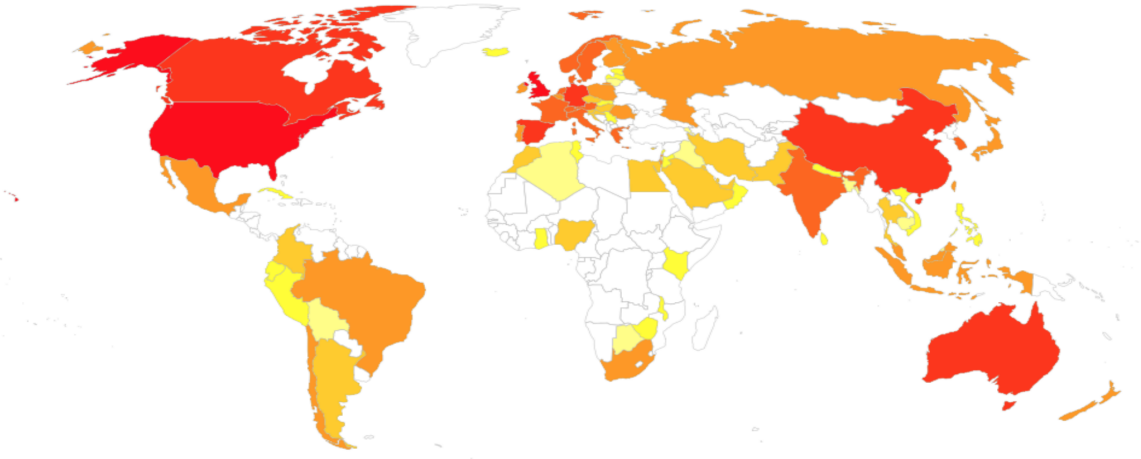

Figure 2.8 Cartographie des publications académiques mentionnant sciences politiques et nouvelles données

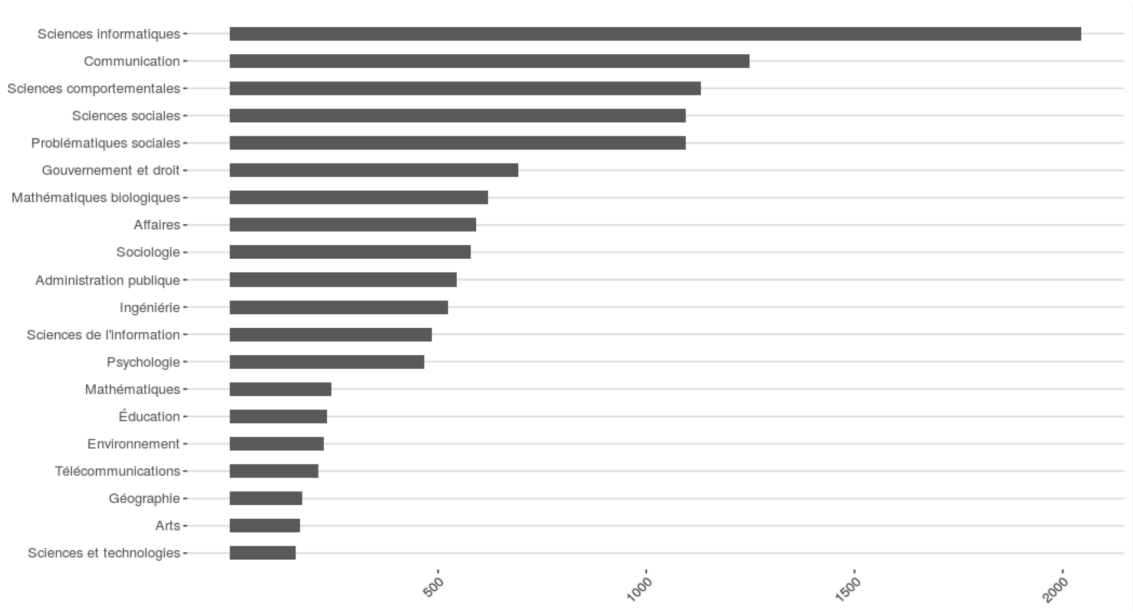

Figure 2.9 Disciplines académiques mentionnant sciences politiques et nouvelles données

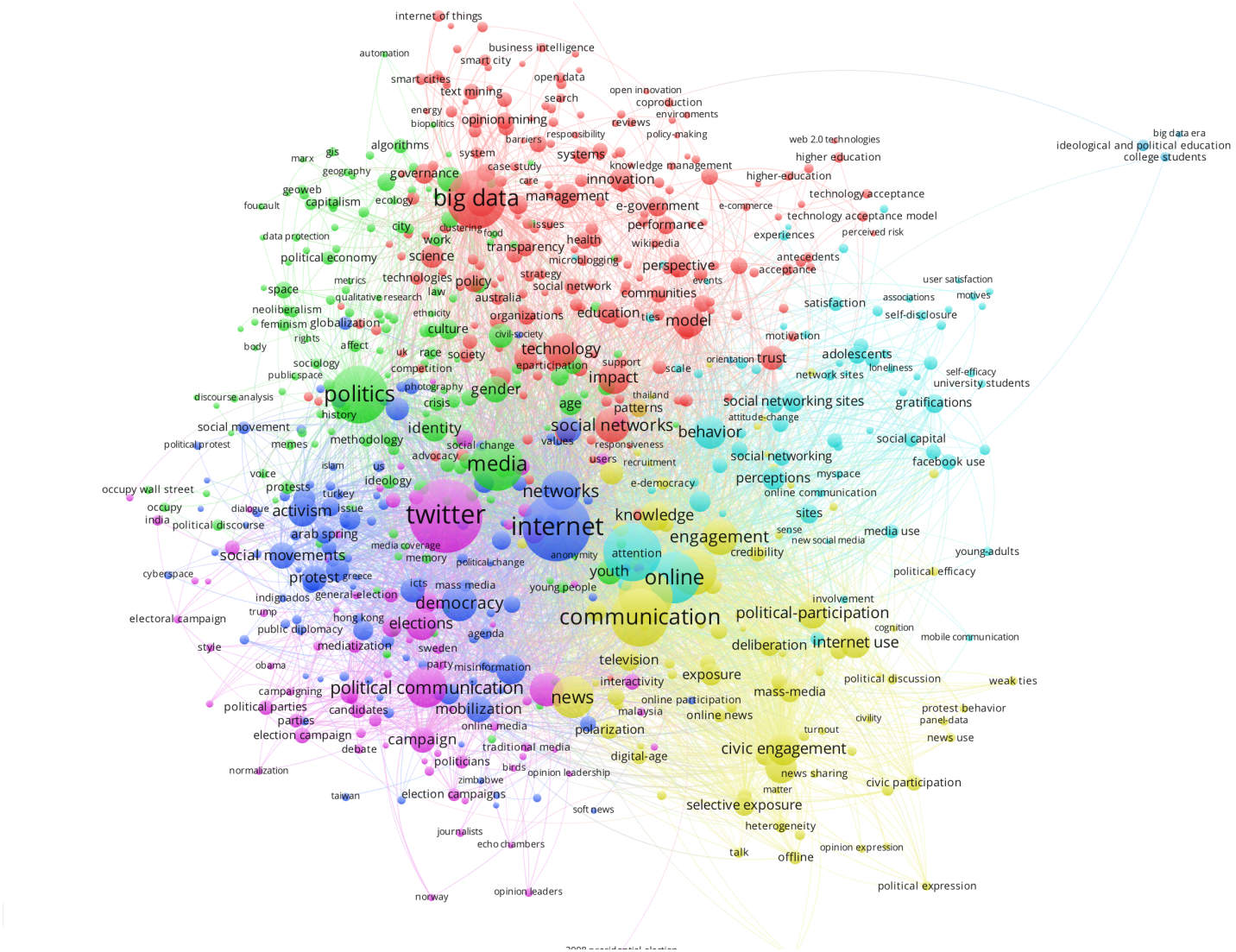

Figure 2.10 Panorama des thématiques de la littérature académique (1998 - 2018)

Tous les mots clefs ont été regroupés par affinité en sept clusters de couleurs différentes. On retrouve les thématiques de la communication, de la participation politique et de l'engagement civique ensemble. La confiance, les études comportementales et les perceptions forment un second groupe. Internet, Twitter, l'activisme et les mouvements sociaux restent fortement associés. Finalement, on remarque que la communication politique, les campagnes électorales, les partis politiques et les élections forment un groupe homogène. D'un point de vue technologique, les données massives, le gouvernement numérique, les sondages d'opinions et la transparence laissent deviner des questionnements en ce qui touche à l'utilisation des nouvelles technologies.

À partir de cette revue de littérature algorithmique, nous dévoilerons plus en détail quatre pans de la littérature. Ils s'articulent à partir d'une approche large dans un premier temps (relations de pouvoir entre institutions, individus et partis politiques) jusqu'à l'étude des phénomènes communicationnels propres aux élections modernes. Ces sections sont les suivantes :

1. les forces et les comportements politiques
2. les médias de masse et la politique
3. la représentativité des populations
4. les élections et les données massives

La première thématique se concentre sur les actions des individus, des partis politiques et des institutions au cours des élections. La deuxième thématique questionne le rôle des médias de masse (télévision, radio, Internet) en politique. La troisième thématique regroupe des études quant à la représentativité de la population lors d'élections, mais aussi par rapport aux sondages et sur les médias sociaux. Finalement, la quatrième thématique observe le rôle des médias sociaux lors des élections des deux dernières décennies.

2.2 Forces et comportements politiques

Comment se justifie l'action politique des individus au sein de la société ? Quelles sont les forces en présence lors d'une élection ? Quelle est l'influence du contexte sur les institutions démocratiques ? À travers cette section seront abordés le rapport des individus à l'information, la question de la représentativité des citoyens dans les décisions gouvernementales, les éléments individuels poussant une personne à aller voter et le rôle des institutions démocratiques.

2.2.1 Information et citoyens

La capacité de comprendre l'information de la part des citoyens est une thématique centrale en sciences politiques. Alors qu'Internet se démocratise depuis le début des années 90, quel rapport à l'information entretiennent les citoyens ? Restent-ils critiques face aux différents éléments d'informations mis à leur disposition ?

En essayant de comprendre si les citoyens basent leurs décisions en comparant leur propre pays aux pays étrangers, [Kayser et Peress, 2012] observent la compréhension de deux variables économiques, soit la croissance et le chômage. Ils trouvent que les électeurs rendent responsables les dirigeants pour la dimension locale de la croissance des pays et non pour la dimension internationale, ne liant les performances/contre-performances des politiciens qu'au niveau local seulement. [Bartels, 2005] met en évidence que les liens de cause à effet et les implications de décisions politiques ne sont pas complètement compris par les individus. Ainsi, il se penche sur les raisons expliquant l'approbation populaire de réductions de taxes survenues lors de la première présidence Bush en 2001. Alors que la plupart des citoyens américains reconnaissent l'existence de disparités sociales, peu ont désapprouvé les réductions de taxes ayant bénéficié aux plus riches.

Un tel phénomène est-il alors à l'avantage des partis politiques ? [Blais *et al.*, 2009] tentent de comprendre si les partis politiques bénéficient systématiquement du manque d'informations de la part des électeurs. Ainsi dans le contexte canadien, les auteurs explorent les raisons pour lesquelles les électeurs les moins informés votent différemment des électeurs les mieux informés. Par exemple, à travers six élections (1988 à 2006), les électeurs les mieux informés votent préférablement pour le Nouveau Parti démocratique (NPD) ou le Parti Conservateur du Canada que pour le Parti Libéral du Canada (PLC). Dans le cas où les électeurs seraient moins informés, alors le PLC se trouverait avantagé, notamment par la notoriété associée au parti politique. Des partis politiques se voient donc favorisés différemment par le manque d'exposition des individus à l'information. Un apport d'information supplémentaire lors d'une

décision à prendre changera toutefois la décision que prendra un individu [Luskin *et al.*, 2002]. En Grande-Bretagne, à la suite d'un sondage délibératif (prise de décision initiale, puis dévoilement d'informations factuelles visant à informer les individus), le changement de positions des individus a été noté dans deux tiers des cas, mettant l'emphase sur l'importance du processus d'apprentissage.

Alors que la littérature académique supporte le fait que la majorité des électeurs ne sont que peu ou pas informés pendant une élection, des raccourcis cognitifs (par exemple avec l'affiliation partisane d'un candidat, l'idéologie d'un candidat, les appuis des candidats, les résultats des sondages ou l'apparence des candidats) sont utilisés par les individus pour orienter leurs choix. Dans une étude de [Lau et Redlawsk, 2001], les auteurs explorent les déterminants individuels et contextuels favorisant l'utilisation de tels raccourcis, et si cela peut affecter la décision politique. De leur étude, trois effets sont à noter : (1) l'utilisation de raccourcis cognitifs peut être substituée au besoin d'accès à de l'information de qualité, (2) l'utilisation de raccourcis cognitifs est au bénéfice d'individus experts ; (3) lorsque le(s) candidat(s) se comporte(nt) de manière non conventionnelle, alors l'utilisation de raccourcis cognitifs est au désavantage des électeurs experts. Mettant l'emphase sur les électeurs dits "naïfs", [Lupia, 1994] questionne la capacité de ces électeurs à compenser leurs lacunes par l'utilisation de raccourcis et ainsi mimer les comportements d'électeurs dits "experts". L'auteur montre qu'il n'est pas nécessaire d'éduquer parfaitement les électeurs, mais plutôt de leur fournir des signaux crédibles avant d'effectuer leur choix de vote.

2.2.2 Représentation de la population au sein des partis politiques

Comment les partis politiques reflètent-ils les positions des électeurs ? [Ezrow *et al.*, 2011] cherchent à savoir si le changement de positions de partis politiques est une réponse au changement de position de l'électeur médian ou plutôt une réponse au changement de position de leur base partisane. À partir de trois hypothèses de travail - (1) le changement de position des partis est causé par le changement de position de l'électeur médian ; (2) le changement de position des partis est causé par un changement de position des partisans du parti et (3) les partis en marge sont plus susceptibles que les partis traditionnels à être influencés par leurs partisans - les auteurs trouvent que les partis de niche ne sont pas sensibles au changement au sein de l'électoral en général, mais plutôt par l'opinion de leurs partisans.

Est-ce que tous les citoyens sont représentés de la même manière ? [Gilens, 2005] observe que les décisions politiques reflètent préférentiellement les classes sociales aisées (par opposition à la classe moyenne et aux personnes à faible revenu). Lorsque l'avis de la population américaine est hétérogène, alors les décisions politiques semblent refléter l'avis des populations à haut

revenus.

La proposition 3 de l'ouvrage de [Downs, 1957] soutient que lorsque deux partis sont en confrontation, leurs positions idéologiques convergeront, notamment vers celles de l'électeur médian. Par extension, est-ce que les candidats convergent vers les politiques de leur parti ou plutôt vers les demandes locales ? À travers une étude ambitieuse reprenant des données de 1874 à 1996 [Ansolabehere *et al.*, 2001], les duos de candidats ne semblent pas converger ; toutefois, le type de compétition aura une certaine importance. En effet, plus le district électoral est polarisé, plus la différence de positions entre les candidats sera prononcée. Historiquement, l'influence du contexte local est aussi variable : des périodes d'extrémisme local ont existé entre les années 40 et 70, mais cette dynamique s'estompe par la suite.

2.2.3 Participation des citoyens

Si les décisions politiques ne reflètent pas l'avis de tous les citoyens, et qu'un vote n'ait qu'une influence marginale quant au résultat final d'une élection, comment donc expliquer le fait que les électeurs se présentent tout de même aux urnes lors d'une élection ? Quels sont les déterminants de la participation électorale ? À travers une étude longitudinale américaine (1969 à 2002), [Aldrich *et al.*, 2011] montrent que la stabilité du contexte des individus (habiter au même endroit plus de dix ans par exemple) en conjonction avec la répétition de l'acte de voter font en sorte que le vote devienne une habitude forte pour les citoyens.

[Blais *et al.*, 2011] mettent en évidence que le sens du devoir et l'intéressement à la politique sont deux facteurs essentiels qui expliquent la propension des individus à aller voter. Toutefois, le modèle de choix rationnels où un individu ne prendra en compte que ses coûts d'accès à l'information et sa fonction d'utilité lors de sa décision montre ses limites (modèle mis en avant dans l'ouvrage de Downs), notamment pour une personne ayant un sens élevé du devoir de voter. En étudiant le rôle du temps, de l'argent et des compétences individuelles dans l'implication électorale, [Brady *et al.*, 1995] montrent que le fait d'aller voter est fortement influencé par l'intérêt politique et très peu par le fait de posséder des ressources financières.

Le changement de position lors d'une élection est aussi lié au degré d'importance des thématiques de campagne et des connaissances de l'électeurs par rapport à ces dernières. [Carsey et Layman, 2006] étudient les circonstances influençant mutuellement l'identification partisane et les préférences par rapport à différents thèmes de société (dans ce cas-ci l'avortement, les dépenses publiques et les questions raciales). Pour qu'un individu puisse changer d'identification partisane, il faut qu'il soit au courant du positionnement des différents partis politiques puis que la thématique abordée soit d'importance pour lui.

De plus, la part des individus dans l'explication du vote se retrouve aussi dans la compréhension de la notion de vote stratégique. En effet, ce dernier pourrait être expliqué par une volonté autre que de rendre son bulletin inutile. Deux éléments influencent ce comportement électoral, soit l'adéquation de l'électeur envers les positions du parti politique visé et ensuite la compréhension des enjeux électoraux [Gschwend, 2006]. [Alvarez *et al.*, 2006] ajoutent le fait que les partis en compétition doivent être au coude à coude dans les intentions de votes et que le parti d'affection de l'électeur ne doit pas obtenir de chance de l'emporter lors du scrutin.

Afin de compléter sur l'influence du contexte local [Ansolabehere *et al.*, 2001], y aurait-il une influence du milieu social dans l'acte de voter ? [Gerber *et al.*, 2008] montrent que non seulement la pression sociale influence positivement la participation électorale, mais aussi que le fait d'exposer les choix de vote d'une personne à ses voisins renforce cet effet.

2.2.4 Institutions démocratiques

Pour terminer, les institutions démocratiques détiennent un rôle déterminant quant aux comportements politiques des différents agents. Par exemple, la France est caractérisée par des élections à deux tours, où les candidats aux élections législatives peuvent accéder au second s'ils remportent un minimum de 12,5% des votes lors du premier tour d'élection. Face à un adversaire commun (le Front National), des alliances électorales peuvent alors voir le jour, comme il a été observé entre le Parti Socialiste et les Verts [Blais et Indridason, 2007].

Comme mentionné précédemment, [Ezrow *et al.*, 2011] se concentrent sur la réponse des partis politiques face à leur base électorale mais aussi face à l'électeur médian. Ainsi, deux forces sont en action : une force centripète où le parti politique répond à des impératifs de modération et une force centrifuge où les positions évoluent vers l'extrémité du spectre politique. En étudiant les élections législatives dans 13 pays différents, [Calvo et Hellwig, 2010] observent que les règles électorales ont une influence sur la position des partis à travers le spectre électoral. Ainsi, les partis traditionnels tendront à adopter une position plus modérée, tandis que les petits partis auront une position plus extrême. Il est intéressant de mettre ce phénomène en perspective avec les récents succès électoraux du camp du Brexit (2016), où 40% des députés du parti au pouvoir, les Conservateurs, a promu la position en faveur du Brexit. En parallèle, la campagne électorale menée par le parti eurosceptique UKIP fut marqué par l'utilisation de fausses informations concernant les contributions européennes et britanniques.

Considérant les différents types de systèmes électoraux, est-ce que le système proportionnel serait en meilleure concordance avec les citoyens que le système majoritaire ? En effet, le système proportionnel permet de refléter dans la constitution des élus les préférences électorales

des électeurs. [Golder et Stramski, 2009] tentent ainsi à mesurer la distance idéologique entre le citoyen médian et les gouvernements élus. Sur 41 élections législatives à travers 24 pays, les auteurs montrent toutefois que la différence de système électoral n'a que peu d'impact sur la congruence des gouvernements (différence entre ce que veulent les électeurs et les positions des gouvernements).

De la même manière, l'importance attribuée aux chefs de partis politiques est différente selon les pays [Holmberg et Oscarsson, 2013]. Aux États-Unis, l'effet du chef de parti est dominant par rapport à l'effet du parti politique. Cette relation est aussi confirmée pour le Canada, l'Australie et la Grande Bretagne. Pour les pays scandinaves, c'est plutôt l'inverse, c'est à dire que l'effet du parti qui est plus important que l'effet du chef de parti.

Tableau 2.1 Références académiques concernant la section forces et comportements politiques

Sujets	Auteurs
Rapport à l'information	Bartels (2005); Blais et al. (2014); Kayser et Peress (2012); Lau et Radlawsk (2001); Lupia (1994); Luskin et al. (2012)
Représentativité électorale	Ansolabehere et al. (2001); Downs (1951); Ezrow et al. (2011); Gilens (2005)
Participation des citoyens	Aldrich et al. (2010); Alvarez et al. (2006); Ansolabehere et al. (2001); Blais et Achen (2018); Blais et al. (2010); Brady et al. (1995); Carsey et Layman (2006); Gerber et al. (2008); Gschwend (2006)
Institutions	Blais et Indridason (2007); Calvo et Hellwig (2011); Ezrow et al. (2011); Golder et Stramski (2010); Holmberg et Oscarsson (2011)

2.3 Médias de masse et politique

L'impact des médias et l'imputabilité des politiciens ont été étudiés à travers différents contextes (types d'élections, pays, époque). La démocratisation des nouvelles technologies de communication, que ce soit avec les journaux, la radio, la télévision ou plus récemment avec Internet, a eu un effet politique indéniable. Ici, nous nous intéressons à connaître l'impact des données massives sur l'espace politique, et en quoi ce type de données change le rôle des médias sur le contexte électoral.

Nous aborderons dans un premier temps l'impact des médias sur l'espace politique tel que mentionné à travers la littérature. Ensuite, après avoir spécifié la nature des données massives en contexte électoral, nous détaillerons le type d'influence que possèdent ces nouvelles données.

2.3.1 Imputabilité électorale et médias "traditionnels"

Le XX^{e} siècle a été le théâtre d'innovations technologiques qui une fois adoptées par la population, ont su influencer le paysage électoral. Les élections, par leurs intervalles réguliers, sont des marqueurs temporels de choix pour comprendre l'influence de la démocratisation des technologies en politique.

L'influence des journaux a été étudiée sous différentes aspects. [Snyder et Strömberg, 2008] utilisent une mesure de congruence pour évaluer la correspondance entre la couverture géographique des journaux et l'étendue géographique des districts électoraux américains. Trois résultats ont été tirés de leurs travaux. (1) Plus la couverture médiatique (via les journaux) correspond à la zone du district électoral, plus les électeurs connaissent leurs représentants. Ces derniers sont alors imputables lors des élections puisque les électeurs connaissent les enjeux électoraux et les décisions ayant été prises par les élus. Dans les districts possédant un niveau de congruence élevé, les auteurs ont trouvé que les transferts d'argent du gouvernement fédéral vers le district électoral sont alors plus importants. Cette notion d'imputabilité électorale est aussi étudiée dans le cadre norvégien par [Bruns et Himmler, 2011].

Les auteurs s'intéressent aussi au niveau de compétition dans le secteur des journaux par rapport à l'efficacité des politiques menées par les élus : l'efficacité (mesurée par le nombre de lois passées selon un budget établi) est alors corrélée positivement avec le nombre de journaux disponibles. La couverture médiatique se doit d'être locale pour que cette relation soit vérifiée, car cela force les politiciens à agir en fonction des attentes des citoyens. Ces derniers sont alors mieux renseignés sur les politiques promues. [Strömberg, 2015] établit une revue systématique par rapport à l'impact des médias de masse sur la politique. Entre autres, il

met en évidence cet enchaînement, couverture médiatique accrue - plus grande imputabilité - meilleure efficacité des politiques, comme reflet du processus de transmission de l'information aux électeurs qui choisissent avec plus de justesse leurs représentants. [Strömberg, 2004] arrivait à la même conclusion une décennie plus tôt par rapport à l'introduction progressive de la radio au sein des foyers américains. Il montre à partir de données au niveau des comtés américains que l'arrivée de la radio a eu un effet important sur la redistribution des aides du programme FERA (Federal Emergency Relief Administration). Pour chaque pourcentage supplémentaire de foyers avec une radio, cette aide a été bonifiée de 0,6%. De plus, chaque pourcentage supplémentaire d'illettrisme est relié à une diminution d'un pourcent de l'aide accordée dans les programmes FERA, notamment par le fait que la population lésée n'est pas informée quant aux implications politiques.

Outre les journaux et la radio, la télévision a été un autre média de masse amplement étudié par la littérature. Tant aux États-Unis que dans les autres pays, son adoption ne s'est pas faite uniformément. Aux États-Unis, des effets différents ont été soulevés. Ainsi, [Gentzkow, 2006] attribue à l'introduction de la télévision un effet négatif sur le taux de participation des électeurs. Entre 25 et 50% de l'augmentation du taux d'abstention est lié à la démocratisation de la télévision dans les foyers américains, ce qui est expliqué par un délaissement des sources traditionnelles d'informations politiques (journaux, radio) au profit d'un média offrant une opportunité de divertissement aux spectateurs. Toutefois, le paysage médiatique américain a fortement évolué au cours des deux dernières décennies, notamment par l'offre de chaînes idéologiquement orientées aux citoyens américains (Fox News et MSNBC pour ne nommer que celles-ci par exemple). Avant d'aborder la notion de médias biaisés, attardons-nous sur le cas italien. Au début des années 2000, la télévision italienne a effectué un virage technologique, passant d'un réseau analogique à un réseau digital, proposant ainsi une plus grande offre médiatique aux électeurs italiens. Cette nouvelle offre a provoqué un changement de perception des politiciens italiens, en particulier par rapport au Président du Conseil Silvio Berlusconi. Alors qu'il était généralement bien perçu, bénéficiant d'un penchant favorable de la part des médias, l'introduction d'une plus grande et plus diversifiée offre médiatique a provoqué une baisse du niveau de soutien du politicien (de 6,5% en moyenne). Environ 20% des utilisateurs de télévision numérique ont changé leur habitude de vote suite à l'introduction de la télévision numérique, entre autre parce que les politiciens s'exposent à des avis critiquant les politiques qu'ils soutiennent [Barone *et al.*, 2015].

[Hopkins et Ladd, 2014] s'intéressent à l'introduction de Fox News et à l'impact d'une chaîne idéologiquement biaisée sur les intentions de vote et à travers les catégories d'électeurs. Grâce à un sondage effectué en 2000 sur 22 595 répondants, les auteurs révèlent que l'introduction d'un média idéologiquement biaisé renforce les convictions partisanes d'une base électorale

déjà encline aux idéologies véhiculées (ici les Républicains) ou au moins arrive à influencer les électeurs Indépendants. De manière générale, l'effet de l'introduction de Fox News a été associé à un vote favorable envers le parti Républicain à hauteur de +1,22% d'intentions de votes dans des municipalités ayant accès à la chaîne de télévision. Cet effet est toutefois différent selon l'identification partisane des individus : +3,7% pour les Indépendants, +2,6% pour les Républicains et +0,34% pour les Démocrates. Si l'on met ce résultat en perspective par rapport aux études précédentes, on peut comprendre que l'introduction d'un tel type de chaîne télévisée renforce la notion d'imputabilité des élus, puisque les citoyens sont interpellés selon leurs convictions partisanes. [DellaVigna et Kaplan, 2007] corroborent cette augmentation d'influence, soulignant qu'entre 1996 et 2000, pour les municipalités ayant accès à Fox News, le Parti Républicain a gagné entre 0,4 et 0,7% de votes.

En termes d'imputabilité, on remarque qu'un changement technologique (l'introduction de la télévision numérique par exemple), est liée à une remise en cause du pouvoir élu. En Russie, cela s'est traduit par un recul de l'appui au gouvernement de 8,9% [Enikolopov *et al.*, 2011]. Aux États-Unis (alors dirigés par le Président démocrate Bill Clinton), c'est un appui favorable envers les Républicains [Hopkins et Ladd, 2014]. La situation en Italie est intéressante, car Silvio Berlusconi fut propriétaire du plus important consortium de médias télévisés avant son arrivée au pouvoir. Lorsqu'il est arrivé à la tête d'une coalition de centre-droit, il a été observé que les médias ont opéré un changement de perspective dans le traitement de l'information en orientant les points de vue vers la droite du spectre politique [Durante et Knight, 2012]. Ce changement s'est aussi reflété dans les préférences de vote des citoyens. En réaction à un contenu médiatique orienté, les sympathisants de gauche se sont tournés vers du contenu médiatique moins biaisé en faveur du gouvernement nouvellement élu.

Toujours en Italie, [Campante *et al.*, 2013] tentent de comprendre les dynamiques électorales qui ont suivi l'arrivée d'Internet. De la même manière que l'introduction de la télévision numérique a provoqué une remise en cause du pouvoir établi, la disponibilité d'Internet a provoqué dans un premier temps un désistement de participation électorale. Toutefois, ce désistement n'est pas causé par un désintéressement politique, mais plutôt par un refus des partis "traditionnels" alors au pouvoir. Après 2005, la formation d'un nouveau parti (*Movimiento Cinque Stelle*, Mouvement 5 Étoiles) dirigé par Beppe Grillo est la matérialisation de l'introduction en Italie du média de masse qu'est Internet.

Les médias de masse ayant un impact sur le paysage politique et l'imputabilité des politiciens, il est nécessaire de comprendre de quelle manière les données massives d'aujourd'hui sont apparentées à ce même type d'influence (ou au contraire, dévient des processus d'influence des médias de masse traditionnels).

2.3.2 Données massives en contexte électoral

En premier lieu, qualifions ce que l'on appelle données massives, et cela en contexte électoral en particulier. Le terme "mégadonnées" ou "Big Data" est un terme flou chargé de significations différentes selon les contextes d'utilisation. Des définitions traditionnelles et implicitement marketing définissent ces données massives comme les 3V [Laney, 2001], 4V, 5V, 6V... (10V ?...) avec de manière générale une emphase sur la variété des types de données, le volume de données produites et la vitesse à laquelle ces données sont générées.

Nous éviterons d'utiliser ce type de définitions restrictives dans cette étude afin d'éviter les approximations. Les mégadonnées sont le résultat d'une combinaison de facteurs. Dans un premier temps, la démocratisation des plateformes sociales du Web permet une production impressionnante de données. À titre d'illustration, ce sont 500 millions de tweets qui sont publiés sur Twitter chaque jour. De plus, la puissance des processeurs informatiques, leur intégration dans des téléphones mobiles ou intelligents et l'accès favorisé à Internet contribuent à la production de données de la part de l'ensemble des populations. Finalement, l'intégration de ces données dans des bases de données "massives" permet le traitement statistique poussé des traces numériques des usagers. Cette convergence est le point de focalisation d'un nouveau type d'économie basé sur la capitalisation des productions numériques des usagers, appelé capitalisme informationnel [Proulx, 2011].

De ces différents facteurs nous pouvons définir les mégadonnées comme des entités ayant pour caractéristiques d'être [Warin *et al.*, 2014] :

1. structurées ou non structurées : cela peut être des photos, des données de géolocalisation, des métadonnées, des données physiologiques...
2. disponibles en grande quantité : les traces numériques sont produites par l'ensemble des individus, des compagnies et des institutions
3. disponibles en temps réel : les téléphones intelligents, les capteurs et les ordinateurs produisent des données en continu
4. le plus souvent longitudinales : les bases de données agrégeant ces données massives permettent une analyse historique

Qu'en est-il en contexte électoral ? L'élection de Barack Obama en 2008 a permis de qualifier les États-Unis de "Nation en réseau" ou "Networked Nation" [Cogburn et Espinoza-Vasquez, 2011]. Depuis ces élections, pas une campagne électorale n'a lieu sans que les réseaux sociaux deviennent un outil de prédilection pour les analystes politiques, pour les politiciens et pour les citoyens. Les partis politiques utilisent des outils technologiques permettant de connaître

les avis des citoyens, mais aussi pour optimiser les stratégies pendant la campagne électorale, notamment pour la gestion des bénévoles lors des porte-à-porte [Liégey *et al.*, 2013]. À titre d'exemple, au cours de la campagne électorale canadienne de 2015, ce sont plus de 3,7 millions de messages qui mentionnèrent un des trois mots-clics politiques utilisés (#polcan, #cdnpoli ou #elxn42) [Sanger et Warin, 2018a].

2.3.3 Influence des données massives par rapport aux médias traditionnels

Il convient de mettre en perspective l'influence de ces données par rapport aux médias traditionnels. Cela permettra d'anticiper les effets potentiels sur le paysage politique (ce qui sera abordé dans les deux parties suivantes). Avant toute analyse, les données massives et le type de communication qui leur est associé se différencient des médias de masse traditionnels par trois éléments caractéristiques.

1) Dans un premier temps, les modes de distribution de l'information sont modifiés. Tandis que les télévisions ou les journaux redistribuent une information vérifiée et calibrée par les comités éditoriaux auprès du public, l'information issue des données massives provient de l'ensemble des individus présents sur Internet. Les sources d'informations sont donc multiples, impliquant de nouvelles voix au débat démocratique tout en posant d'autres problématiques, notamment en ce qui concerne la crédibilité des messages publiés [Castillo *et al.*, 2011, Flanagin et Metzger, 2000].

2) Ensuite, les barrières à l'émission de messages et à l'accessibilité de ces messages tendent à disparaître. La publication d'un message politique incombait auparavant aux journaux, télévisions, radio et politiciens. Ceci pose la question des conflits d'intérêts, personnifié par la relation entre l'empire médiatique italien Mediaset (propriété de Silvio Berlusconi) et la direction du pays par ce dernier. Depuis le début des années 2000, tout individu est capable d'émettre une opinion (qui peut être politique) qui sera retransmise via Internet, notamment par les médias sociaux. En ce sens, l'effet des mégadonnées sur la vie politique s'apparentera aux effets de l'introduction de l'Internet comme mentionné dans la littérature et à travers l'article de [Campante *et al.*, 2013].

3) Finalement, de par leur instantanéité et la profondeur d'information qu'elles fournissent, il serait raisonnable de penser que les données massives aient un rôle d'anticipation politique. Les études sur la capacité d'anticipation de résultats électoraux sont nombreuses, notamment sur l'utilisation de Twitter [Bermingham et Smeaton, 2011, Livne *et al.*, 2011, Metaxas *et al.*, 2011, Tumasjan *et al.*, 2010]. Nous explorerons plus en détail cet élément dans la prochaine section de la revue de littérature, tout en mettant en avant les limites méthodologiques liées à de telles prédictions.

Ainsi, pour comparer l'influence des données massives avec les autres médias de masse en termes d'imputabilité politique, il est important de comprendre que les données massives apportent aux citoyens l'information nécessaire pour qu'ils puissent évaluer les actions de leurs représentants. Toutefois, ce que des citoyens peuvent gagner en moyen de communication ou d'organisation, des partis politiques obtiennent en contrepartie une information de meilleure qualité concernant les intentions de vote des individus.

Tableau 2.2 Références académiques concernant la section médias de masse et politique

Sujets	Auteurs
Imputabilité électorale et médias "traditionnels"	Barone (2015); Bruns et Himmler (2011); Campante et al. (2013); Della Vigna et Kaplan (2007); Durante (2012); Enikolopov et al. (2011); Gentzkow (2006); Hopkins et al. (2014); Snyder et al. (2008); Strömberg (2004); Strömberg (2015)
Données massives en contexte électoral	Laney (2001); Proulx (2011); Warin et al. (2014); Cogburn (2011); Liégey et al. (2013); Sanger et Warin (2018)
Influence des mégadonnées par rapport aux médias traditionnels	Flanagin et Metzger (2000); Castillo et al. (2011); Campante (2013); Tumasjan et al. (2010); Matexas et al. (2011); Livne et al. (2011); Bermingham et al. (2011)

2.4 Représentativité des populations : public, sondages et réseaux sociaux

Il est ici question de la représentativité des échantillons lors des analyses d'informations issues des médias sociaux. Est-ce que les sondages sont représentatifs de la population ? À titre d'exemple et sur trois continents différents, ni le Brexit de 2015 ni les élections de Muhammdu Buhari (Nigeria, 2015) et de Donald Trump (États-Unis, 2016) ne purent être prédits avant le dévoilement des résultats. Est-ce causé par une dissonance entre les populations qui se déplacent lors des scrutins électoraux et les populations ciblées par les instituts de sondages ? Est-ce que l'utilisation des données massives non structurées provenant des médias sociaux pourrait compléter ces sondages ? Si tel est le cas, qu'en est-il de la représentativité de la population sur des réseaux sociaux tel que Twitter ou Facebook ? Finalement, comme caractériser les informations issues de Twitter, notamment entre les différents types de messages (géolocalisés par exemple).

2.4.1 Prédiction électorale

En termes méthodologiques, pourquoi se concentrer sur les hashtags (mots-clics) pour étudier les élections ? Tout d'abord, les différents hashtags utilisés lors d'une campagne électorale sont mis en avant par les stations de télévisions, les sites web ainsi que les partis politiques et les politiciens. Cette promotion incite les citoyens et acteurs politiques à canaliser leurs discussions vers un lieu virtuel dédié.

Comme mentionné par [Bruns et Moe, 2014], trois niveaux de communications caractérisent Twitter. Au niveau micro, l'interpellation des utilisateurs permet l'établissement d'une communication interpersonnelle et éphémère. Au niveau meso, le réseau de comptes d'utilisateurs se suivant (followers-followees) peut être analysé et ainsi constituer l'audience potentielle d'un utilisateur. Toutefois, les auteurs montrent qu'au niveau macro, les hashtags permettent la formation d'"assemblées ad hoc", c'est à dire un regroupement d'individus spécifiquement dédiés à une cause en particulier, comme lors des campagnes politiques par exemple.

Ainsi, un mot-clic au sein d'un message publié sur Twitter dénote trois caractéristiques :

- le message a la possibilité de rejoindre un public plus vaste que le réseau de followers de l'émetteur initial ;
- il permet la coordination de messages concernant un sujet en particulier ;
- il exprime une volonté de l'utilisateur de prendre part à un processus de communication plus large.

Pour ces raisons, la collecte de messages lors d'une élection sous l'égide d'un mot-clic offre l'opportunité d'observer des dynamiques de transmission de l'information qui seraient diffi-

cilement perceptibles à travers un sondage. Rapidement, les chercheurs ont essayé de prédire les résultats d'élections. Ces prédictions peuvent être appréciées à travers le taux moyen d'erreur absolue de prédictions électorales. Ainsi, des résultats obtenus à partir de l'analyse de messages sur Twitter se trouvent dans les marges d'erreur habituelles des sondages (voir [Bermingham et Smeaton, 2011, Metaxas *et al.*, 2011, Tumasjan *et al.*, 2010], avec un taux d'erreur absolue de 1,65%, 1,1% et 5,85% respectivement).

2.4.2 Représentativité des échantillons

À l'instar des sondages basés sur les données provenant de médias sociaux, des erreurs de mesure subsistent dans l'utilisation de sondages "traditionnels". En effet, [Ansolabehere *et al.*, 2008] se penchent sur la présence d'erreurs de mesure dans les sondages et proposent une méthode afin de limiter l'effet de telles erreurs. En posant le plus de questions possibles sur des thématiques particulières au cours d'un sondage, alors les erreurs de mesures se trouvent réduites. Ce phénomène permet d'approximer les convictions des électeurs, et ainsi leur future décision de vote.

Concernant les données publiées sur les médias sociaux, dans quelle mesure les utilisateurs de médias sociaux sont-ils représentatifs de la population en générale, et de la population qui vote en particulier ? [Mellon et Prosser, 2017] étudient comment les échantillons d'individus sur Twitter, Facebook et à travers la population diffèrent en termes de démographie, d'attitudes politiques et de comportements politiques. Les trois types de population ne sont pas interchangeables, soulevant la nécessité de cerner les études utilisant ce type de données. En effet, un rebalancement des échantillons est nécessaire afin d'extrapoler des résultats entre des échantillons issus des médias sociaux vers les comportements de la population en général (contrôlant pour l'âge, le genre et l'éducation). Les populations sur Twitter et Facebook sont plus jeunes et plus éduquées que leur contrepartie dans la population, tandis que la représentativité féminine est plus forte sur Facebook et à l'inverse, la représentativité masculine est plus élevée sur Twitter. Finalement, les individus présents sur les réseaux sociaux, alors plus volubiles que la population en général, votent moins.

À travers une étude réalisée en 2011 (alors que Twitter était en opération depuis seulement 5 ans), [Mislove *et al.*, 2011] se penchent sur les caractéristiques des individus mentionnant leur localisation aux États-Unis. Leur échantillon équivaut à environ 1% de la population américaine et n'est que peu représentatif, ayant à nouveau des biais en faveur des hommes, des régions fortement urbanisées, et surexposant les minorités ethniques. Il faut toutefois noter qu'entre 2007 et 2009, la proportion d'hommes par rapport aux femmes a diminué, suggérant un rapprochement dû à l'adoption de Twitter par la population dans son ensemble.

Plus de dix ans après sa mise en service, où en est la représentativité des individus sur le réseau social ? De même, les études qui emploient des données issues de Twitter devraient mettre en perspective le type de représentativité électorale. Dans un système proportionnel où un représentant est élu directement, alors tenir compte des tweets se rapprocherait plus de la réalité que pour des systèmes où la dimension géographique est importante (pour l'élection de députés par exemple) où le rebalancement des échantillons est nécessaire, entre autres par l'ancrage géographique des données.

En écho à cette dernière étude, [Malik *et al.*, 2015] cherchent à évaluer la représentativité des messages géolocalisés émis aux États-Unis. En effet, Mislove et al. (2011) ne considèrent que les régions reportées par les utilisateurs, induisant potentiellement des erreurs d'échantillonnage. De plus, cette étude rapporte les messages à l'échelle des états américains. À l'inverse, [Malik *et al.*, 2015] utilisent les coordonnées de géolocalisation, permettant une approche plus fine de chaque message (au dix millième de degré de longitude et de latitude). Ils trouvent que les messages géolocalisés ne sont pas représentatifs de la population américaine après avoir associé chaque message à un "block" (il existe plus de 11 millions d'unités géographiques - block - aux États-Unis lors du recensement de 2011). Plus précisément, la population émettant des messages géolocalisée présente les caractéristiques démographiques suivantes : c'est une population ayant un revenu médian plus élevé que la moyenne, étant plus jeune et plus urbaine que la moyenne ; elle est surreprésentée dans les régions côtières.

De manière générale, [Hecht et Stephens, 2014] mettent en évidence que les réseaux sociaux reposant sur des systèmes volontaires de géolocalisation déforment la représentativité des populations rurales et urbaines. En effet, les régions urbaines américaines sont surreprésentées à travers les échantillons de tweets géolocalisés, à raison de 5,3 fois plus de messages par utilisateur, ou entre 2,7 et 3,5 fois plus d'utilisateurs par circonscription géographique. Par rapport à Twitter en particulier, l'entreprise permet la collecte d'informations à plusieurs niveaux (voir la section méthodologie de la thèse). En effet, une interface de programmation (API) permet d'accéder gratuitement à environ 1% de l'ensemble des messages publiés sur la plateforme (Streaming API). Afin de récupérer l'ensemble de l'information, il est nécessaire de se connecter à un service payant et exigeant technologiquement, le Firehose API. [Morstatter *et al.*, 2013] ont évalué les différences structurelles entre ces deux ensembles de messages pour savoir si l'utilisation des messages issus du Streaming API est représentative de l'ensemble des messages publiés sur la plateforme. L'identification des mots clics les plus populaires n'est pas parfaite en utilisant le Streaming API ; toutefois, près de 90% des messages géolocalisés s'y retrouvent. Ainsi, collecter les messages géolocalisés à l'aide du Streaming API permettrait de capturer près de l'ensemble des messages géolocalisés de la plateforme.

Dans le cadre des élections italiennes de 2013, [Vaccari *et al.*, 2013] étudient l'influence des messages publiés en ligne et le potentiel effet de contagion dans les conversations hors ligne. Ils trouvent que le comportement des individus sur Twitter ne diffère pas de ceux de la vie quotidienne. Ainsi, plus les individus interviennent dans des conversations politiques sur Twitter, plus ils sont à même de discuter politique hors ligne et ainsi faire circuler de l'information provenant d'Internet. L'information en ligne peut alors influencer des individus à deux niveaux : en ligne, puis dans un second temps hors ligne.

Toutefois, ils remarquent que les individus sur Twitter ne sont pas parfaitement représentatifs de la population. En effet, les utilisateurs concernés par leur étude sont moins religieux, plus éduqués, plus jeunes, plus à même d'être employés ou aux études, de préférence de sexe masculin, que la population italienne en général. À l'inverse, des ensembles de la population électorale ne sont que peu représentés (personnes de plus de 55 ans, alors qu'ils constituent un tiers de l'électorat). De la même manière, en comparant les élections espagnoles de 2013 et les élections américaines de 2012, [Barberá et Rivero, 2015] se penchent sur les habitudes d'utilisation de Twitter de la population. Ils veulent connaître si la population est adéquatement représentée sur la plateforme ou non. Cinq traits caractéristiques émergent de leur analyse : les participants aux conversations politiques sont plutôt des hommes, situés en région urbaine et ayant de fortes positions idéologiques ; ces utilisateurs participent aussi plus facilement s'ils supportent un parti politique, les individus conservateurs plus que les individus libéraux. Ces résultats mettent en avant le fait que la participation sur Twitter n'est pas homogène, et que ces biais d'échantillonnage doivent être pris en compte.

2.4.3 Limites méthodologiques à considérer

La précédente littérature met en avant le fait que la population en général, et votant en particulier, n'est pas parfaitement représentée sur Twitter. [Filho *et al.*, 2015] mentionnent la nécessité de rebalancer les échantillons afin d'éviter de potentielles erreurs de mesure lors de prédictions électorales.

Tandis qu'un pan de la littérature propose des avenues de recherche encourageantes par rapport aux prédictions électorales, le manque de processus de recherche reproductibles a été souligné [Chung et Mustafaraj, 2011]. Ainsi, la capacité de prédiction de Twitter semble être moindre que celle de Facebook [Cameron *et al.*, 2013]. Des hauts taux d'erreurs de prédiction sont aussi obtenus, notamment par [Choy *et al.*, 2012] concernant les élections de Singapour en 2011 où le taux d'erreur moyen de prédiction était de 6,06%. En analysant 234 000 messages concernant les élections sénatoriales américaines de 2010, [Gayo Avello *et al.*, 2011] obtiennent des marges d'erreurs de 17,1%, qui tombent à 7,6% en prenant en compte

le sentiment associé à chaque message publié sur Twitter.

Parmi les faiblesses mises en lumière par les auteurs, [Gayo-Avello, 2012a] souligne que la littérature académique présente un biais favorable aux études réussissant la prédiction juste des élections. De plus, deux phénomènes ne semblent pouvoir être capturés, soit les rumeurs et la propagande d'un côté, et la diversité démographique de l'autre. Dans le premier cas, des études plus récentes essayeront de mettre en lumière de tels phénomènes, notamment après les élections américaines de 2016.

En contexte canadien, [Small *et al.*, 2014] se penchent sur la propension des Canadiens à participer sur les plateformes sociales. Alors que la grande majorité des élus au Parlement sont inscrits sur Twitter (près de 80%), seulement une fraction des citoyens se trouve sur le réseau social (3.9% selon un recensement de 2014 au Canada, avec 3,1% de la population ayant écrit un tweet). Ces chiffres suggèrent que la participation canadienne reste faible. Toutefois, est-ce que cela doit remettre en cause l'utilisation de telles données en politique ? Trois raisons montrent le contraire.

1) L'étude de [Small *et al.*, 2014] a été réalisée hors période électorale. Pendant une campagne électorale, les médias et journaux télévisés dédient une couverture médiatique plus importante au contenu politique qu'en temps normal. À titre d'illustration, les pancartes électorales canadiennes sont disséminées le long des routes, offrant l'opportunité à tout citoyen de parler politique dans son horaire quotidien. 2) Un reproche qui peut être fait à la plateforme Twitter est qu'une faible portion d'individus produit une large quantité de contenu. À l'inverse, une large foule silencieuse est présente sur le réseau social. Toutefois, ce phénomène mime les interactions réelles entre individus, où les avis politiques ne sont pas partagés par tous et où certains occupent plus d'espace que d'autres [Barberá et Rivero, 2015, Vaccari *et al.*, 2013]. 3) Comme mentionné précédemment dans la revue de littérature, si un individu ne participe pas aux discussions sur Twitter, il peut toutefois rester influencé par un message ayant été émis par une autre personne [Vaccari *et al.*, 2013].

Tableau 2.3 Références académiques concernant la section représentativité des populations

Sujets	Auteurs
Prédiction électorale	Bruns et al. (2014); Bermingham et Smeaton (2011); Metaxas et al. (2011); Tumasjan (2010)
Représentativité des échantillons	Ansolabehere et al. (2008); Barbera et al. (2014); Hetch et Stephens (2014); Malik et al. (2015); Mellon et Prosser (2017); Mislove et al. (2011); Morstatter et al. (2013); Vaccari et al. (2013)
Limites méthodologiques	Barbera et Rivero (2015); Cameron et al. (2013); Choy et al. (2011); Chung et Mustafaraj (2011); Filho et al. (2015); Gayo-Avello (2011); Gayo-Avello (2012); Small et al. (2014); Vaccari et al. (2013)

2.5 Élections et données massives au XXI^{e} siècle

Ce qui est appelé "gouvernement numérique" est défini comme "l'introduction, la gestion et l'utilisation de technologies de l'information dans le secteur public et les relations avec les citoyens, les entreprises, les ONG et les autres organisations" [Lips, 2014]. Deux visions s'affrontent alors concernant la relation entre technologie et gouvernement : d'un côté, la technologie est un vecteur de transformation du gouvernement, permettant la constitution d'un gouvernement virtuel, allant de la démocratie directe jusqu'à la surveillance de masse. À l'inverse, la technologie peut être perçue comme accompagnant la transformation civique, en positionnant le citoyen au centre de l'action gouvernementale ou en dévoilant ses données ("gouvernement ouvert"). [Lips, 2014] souligne le rôle déterminant des dirigeants politiques pour accompagner le gouvernement vers une transformation digitale ainsi qu'une profonde compréhension de la gestion des problématiques publiques afin de prendre en compte les citoyens délaissés par la digitalisation (jeunes citoyens, personnes âgées, familles au revenu faible, personnes atteintes de maladies...).

À la lumière de cette dualité entre technologie et gouvernement, quels phénomènes peuvent-être étudiés par le prisme des données massives en contexte électoral ? Deux thématiques seront abordées, notamment les dynamiques de campagne et la polarisation des individus sur ces plateformes.

2.5.1 Dynamiques de campagne

Les études électorales utilisant des données issues de Twitter se basent sur une stratégie de cadrage (framing strategy). Certes, l'ensemble de l'information n'est pas pris en compte,

mais cette méthodologie reste standard à travers la littérature. [Elff, 2013] utilise une telle stratégie afin de recréer le positionnement des partis politiques à partir de textes politiques codifiés. La véracité des indicateurs ne peut qu'être aussi précise que les catégories utilisées. C'est une limitation de telles recherches, puisque des catégories ne reflètent pas parfaitement les plateformes électorales [Laver et Garry, 2000]. Dans le domaine des communications, l'utilisation d'une stratégie de cadrage est issue d'un processus inductif basé sur l'identification de thématiques principales [Chong et Druckman, 2007]. C'est un socle sur lequel peuvent s'effectuer l'analyse de données textuelles comme des articles de journaux à travers le temps et les sources d'informations [Monroe *et al.*, 2008]. L'amélioration des thématiques identifiées peut donc favoriser le degré de précision des phénomènes étudiés, comme dans le cadre des études utilisant des données provenant de Twitter.

Traditionnellement, le succès électoral d'un candidat ou d'un parti est étroitement lié à ses dépenses électorales. [Johnston *et al.*, 2004] mesurent l'effet des nouvelles et des publicités sur les intentions de vote des citoyens américains lors des élections présidentielles de 2000. La dynamique de campagne entre Georges Bush et Al Gore est alors différente. Le fait d'avoir investi plus de ressources au début de la campagne de la part de G. Bush a permis de juguler la remontée observée dans les intentions de vote envers Al Gore, permettant au candidat Bush de remporter l'élection à travers les Grands Électeurs.

Parmi les événements d'une campagne électorale figurent aussi les débats télévisés. Pouvoir avoir accès à une source de données en temps réel des réactions des téléspectateurs permettrait de mieux cerner la portée des propositions politiques des politiciens. Ainsi dès 2008, les débats télévisés entre politiciens canadiens ont été analysés sur Twitter [Elmer, 2013]. De la même manière, lors des élections norvégiennes de 2011, la plateforme de Twitter a permis de confirmer que les discussions reflétaient les propositions énoncées par les différents candidats [Karlsen, 2011]. Le support aux différents partis mais aussi la critique des débats ont pu être mis en évidence. Toujours en contexte scandinave (haut niveau de participation et taux de pénétration d'Internet élevé), [Weller *et al.*, 2014] notent que 20% des messages publiés lors d'une campagne électorale proviennent de la dernière journée menant au vote. De plus, ils observent que Twitter a été utilisé lors de trois élections comme un mégaphone pouvant amplifier les débats de société existants.

En se concentrant sur 415 000 messages mentionnant le mot-clic #ausvote lors des élections australiennes de 2011, [Bruns et Burgess, 2011] montrent que 35% des messages publiés sur Twitter sont des retransmissions de messages (retweet, RT), tandis que 20% de ces messages s'adressent directement à d'autres usagers. Les auteurs ont aussi montré que les réseaux d'individus s'adressant mutuellement ou retransmettant leurs messages diffèrent de par leurs

caractéristiques [Burgess et Bruns, 2012]. En Corée du Sud, [Song *et al.*, 2014] associent les candidats à l'élection présidentielle à certaines thématiques de campagne en fonction des termes les plus fréquemment mentionnés.

Les discussions sur Twitter semblent suggérer que le degré d'inclusion et d'interaction est plus poussé qu'à travers les nouvelles politiques des médias traditionnels comme la télévision ou les journaux [Ausserhofer et Maireder, 2013]. Ainsi en 2013 en Autriche, les thématiques discutées sur Twitter étaient différentes de celles présentées à travers les médias. Tandis que les organes de presse traditionnels faisaient intervenir experts et dévoilaient de l'information factuelle face à certains évènements, les discussions présentes sur Twitter laissaient place à l'interprétation des faits et la recherche et le partage d'informations supplémentaires.

Finalement, une des dynamiques mises en évidence au cours d'élections depuis 2015 est la présence de contenus fortement automatisés sur les plateformes en ligne. Twitter permet le développement de robots à travers la création d'applications automatiques. Dans un ouvrage datant de 2014, [Mowbray, 2014] se penche sur les phénotypes des comptes automatisés, nommés en anglais "bots". Elle présente plusieurs types de bots dont les fonctionnalités sont les suivantes :

— *Marketing bots* : promouvoir une institution, entreprise ;
— *Useful bots* : recherche d'informations ;
— *Entertaining bots* : interaction avec utilisateurs ;
— *Internet of toasters* : "internet of things", des objets connectés qui tweetent ;
— *Antisocial bots* : utilisés en politique pour nuire à la réputation des politiciens ou attirer l'attention sur un sujet en particulier.

De plus, elle souligne l'importance de méthodes permettant de connaître la vraie nature d'un compte Twitter (automatisé ou personnel). Comme mentionné par [Mowbray, 2014], "une campagne utilisant de multiples comptes automatisés pourrait causer de larges distorsions". Dans le cadre électoral, la tenue des élections récentes (vote du Brexit, élections américaines de 2016, propagation de fausses informations) viendra soutenir son affirmation.

2.5.2 Polarisation des individus

Les fonctionnalités de la plateforme Twitter permettent donc aux individus d'échanger entre eux. De plus, les interactions entre individus se retranscrivent à travers les métadonnées associées à chaque message publié. Un des phénomènes remarqués sur les plateformes sociales est la présence de chambres d'échos [Barberá *et al.*, 2015]. Ce phénomène est d'autant plus marqué que les firmes technologiques ont une influence sur les modes de communications

politiques, illustré notamment lors des élections américaines de 2016 [Kreiss et McGregor, 2018].

[Passmann *et al.*, 2014] analysent trois types d'interactions utilisées sur Twitter : les retweets, les favoris et les réponses aux utilisateurs. Ils montrent que les premiers sont structurés autour de valeurs communes, telles que l'affiliation partisane, tandis que les derniers n'en sont pas affectés. Les auteurs nomment ce concept "cartel de retweets", puisque les échanges s'ancrent dans une culture de dons engendrant des dettes mutuelles entre utilisateurs, à l'image de la société des Tiv au Nigeria, comme décrit dans l'ouvrage de [Graeber, 2011]. Les réseaux d'interactions entre utilisateurs sur Twitter présentent des différences significatives, puisque les retweets mettent en évidence des regroupements (clusters) d'individus.

Les analyses de réseaux permettent de mettre en évidence de tels comportements. [Larsson et Moe, 2013] observent les utilisateurs de Twitter les plus volubiles lors des élections danoises de 2011 à travers le mot-clic #fv11. Ils se penchent notamment sur les types d'individus (citoyens, experts, médias et politiciens) et s'intéressent à l'implication des citoyens dans le débat public. Ils utilisent entre autres le logiciel Gephi afin de représenter les différents types de réseaux de transmissions de l'information. Durant les élections américaines de mi-mandat en 2010, [Conover *et al.*, 2011] se basent sur 250 000 tweets afin de montrer la polarisation des utilisateurs de la plateforme. Ils différencient deux types de réseaux, le réseau de retweets et le réseau de mentions. Dans le premier cas, les utilisateurs restent fortement polarisés, retransmettant les messages de personnes avec lesquelles ils s'accordent politiquement. Dans le second cas, des liens entre différents groupes ont été relevés.

Ce type de communication différenciée a été observée dans le cadre canadien lors des élections de 2011 [Gruzd et Roy, 2014]. À partir de 5 918 messages provenant de 1 492 utilisateurs, les auteurs montrent que malgré une certaine polarisation des individus, des connections inter-idéologiques peuvent être établies. Par exemple, des utilisateurs ayant des affinités pour les partis situés plutôt à la gauche du spectre politique (Parti Libéral du Canada, Nouveau Parti démocratique, Parti Vert du Canada) échangent préférentiellement entre eux, tandis que les supporteurs du Parti Conservateur du Canada adoptent une communication plus frontale et sarcastique. À travers l'utilisation du mot-clic #exln41 (41è élection fédérale canadienne), les auteurs notent que les individus politisés ont conscience de la présence d'autres points de vue que le leur.

Tableau 2.4 Références académiques concernant la section élections et données massives

Sujets	Auteurs
Technologie et gouvernement	Lips (2014)
Dynamiques de campagne	Bruns et Burgess (2011); Chong et Druckman (2007); Elff (2013); Elmer (2013); Johnston et al. (2004); Laver et Garry (2000); Maireder et al. (2014); Monroe et al. (2008); Mowbray (2014); Song et al. (2014); Weller et al. (2014)
Polarisation des individus	Barbera (2014); Conover et al. (2011); Graeber (2011); Gruzd et Roy (2014); Kreiss et McGregor (2017); Larsson et Moe (2013); Paßmann et al. (2014)

2.6 Nouvelles questions de recherche issues de la littérature

Le contexte politique actuel est complexe ; les données massives et les nouvelles techniques d'analyses offrent une opportunité d'explorer de nouvelles questions de recherche. Comme mentionné en introduction, on note une diminution de la participation électorale et en même temps une montée du populisme à travers le monde.

Une des nouvelles questions de recherche émergeant de la littérature est de comprendre comment mesurer le populisme à l'ère des médias sociaux. [Reynié, 2013] définit dans son ouvrage les nouveaux populismes et explique trois raisons à la montée du populisme en Europe : le vieillissement démographique, la mondialisation et la montée religieuse à travers plusieurs pays. Il explique que des partis dits populistes peuvent bénéficier d'un effet positif en attisant la fibre du populisme patrimonial. Ce concept désigne "une offre politique nouvelle conçue pour tirer un profit politique d'une double inquiétude désormais installée chez les Européens concernant non seulement leur patrimoine matériel, ou leur niveau de vie, mais aussi leur patrimoine culturel, c'est à dire leur mode de vie" (p. 37). Qu'en est-il de l'offre politique des différents partis à travers l'Europe par exemple ?

À la lumière de la littérature empirique existante, il est nécessaire de recadrer l'usage des mégadonnées dans le paysage politique au XXI^{e} siècle. Plusieurs phénomènes sont alors à anticiper.

Ainsi, le taux de participation aux différentes élections devrait augmenter, comme il a été observé en Italie. Par une plus grande implication politique des individus, par un accès jusqu'alors incomparable à une information objective, et par une plus grande représentativité

de l'avis des citoyens (l'exemple des *Meetup* tenus en Italie pour le début du Mouvement 5 Étoiles corrobore cette intuition), les électeurs devraient être plus nombreux à voter lors des prochaines élections. Ce phénomène peut toutefois être nuancé puisque de nouveaux phénomènes sont depuis 2016 observés. Que ce soit lors du référendum du Brexit en 2016, lors des élections américaines de 2016 ou lors de la campagne électorale brésilienne sur Whatsapp, du contenu automatisé fut livré sur les médias sociaux, en ciblant spécifiquement des groupes d'électeurs grâce aux modèles de diffusion de publicités des plateformes numériques. En utilisant la configuration de ces plateformes, il semble alors possible que des groupes de la population peuvent être influencés, et amenés par exemple à ne plus s'exprimer lors d'un scrutin électoral ou au contraire à se mobiliser davantage.

Si les partis politiques traditionnels n'incorporent pas la prise en compte de ces nouvelles données dans leurs processus de décision, alors de nouveaux partis pourraient émerger (Mouvement 5 Étoiles, *Partido de la Red*, mouvement Democratech en France). Dans tous les cas de figure, la prise en compte de l'avis des électeurs sera certainement modifiée. À l'heure actuelle, les citoyens se prononcent à intervalles réguliers pour élire leurs représentants (pour des cycles de 4, 5 ou 7 ans), ou par le biais de référendum (comme en Suisse). Cette structure répond à de nombreuses contraintes, notamment l'organisation d'un recueil systématique de l'information. Toutefois, en raison du niveau de connectivité des individus au XXI^{e} siècle, il est fortement probable que les décisions politiques seront influencées par ces mégadonnées.

La question de la représentativité des individus sur les médias sociaux ou sur Internet a pu poser des problèmes méthodologiques concernant la représentativité démographique. L'accessibilité aux technologies augmente à travers la population, mais son usage ne s'est pas répandu de la même manière entre les différents groupes de la société. Par exemple, la distribution des messages publiés sur des plateformes comme Twitter montre une forte proportion de messages émis par une fraction d'utilisateurs. Les différentes couches de la population n'ont pas non plus la même aisance pour utiliser de tels outils [Pasquier, 2018]. Malgré les biais entre les différents échantillons de la population représentés sur les réseaux sociaux, ces échantillons peuvent être recalibrés de la même manière que le sont les sondages téléphoniques. En effet, la distribution et les habitudes de la population présente sur les médias sociaux peut être connue à travers des rapports publiés à intervalles réguliers par les plateformes en ligne et aux États-Unis par l'Institut Pew.

Concernant les relations entre dirigeant·e·s politiques et les électeurs, deux dynamiques peuvent être anticipées. (1) Tout d'abord, l'avantage du candidat sortant n'est plus assuré, puisque les électeurs ont accès à une quantité d'information jusqu'alors inégalée pour suivre l'application et l'efficacité des propositions électorales de leurs représentants. La facilité d'or-

ganisation de mouvements sociaux à travers les plateformes numériques favorisera l'activisme citoyen. La coordination des regroupements lors des printemps arabes est une de ces manifestations dès 2011, et plusieurs pays bloquent l'accès aux médias sociaux ou à divers services web lors de la tenue de scrutins électoraux. (2) À l'inverse, les dirigeants n'auront jamais eu autant d'informations pour calibrer leur message politique. Les réactions des électeurs en temps réel vont permettre de mesurer l'impact de propositions électorales auprès de la population, devenant des leviers stratégiques lors des campagnes électorales.

En résumé, les données massives ont une influence certaine sur le paysage électoral actuel, de par les approches et les stratégies des différents acteurs, mais aussi par la prise en compte de l'avis des citoyens et le niveau d'information qui seront à leur disposition. Toutefois, cette dynamique de transformation du paysage électoral ne semble pas révolutionnaire, puisqu'elle s'inscrit à travers l'influence des médias de masse en contexte politique. Néanmoins, il faut souligner que ce type de révolution technologique permet l'émergence d'un mode de communication distribué en réseau et non plus monopolisé par les médias dits traditionnels. Cela permet l'ouverture du champ de recherche des comportements politiques à de nouvelles données et à de nouvelles méthodologies pour comprendre les dynamiques électorales.

CHAPITRE 3 DÉMARCHE DE L'ENSEMBLE DU TRAVAIL ET ORGANISATION DE LA RECHERCHE

3.1 Question de recherche et hypothèses

De la revue de littérature précédemment exposée ont été tirées plusieurs opportunités de recherche potentielles. Ces axes de recherche peuvent être reformulés en une question principale qui fait l'objet de cette thèse doctorale :

Comment les données massives et la science des données peuvent être utilisées pour comprendre les processus démocratiques à l'ère d'Internet ?

Le terme "données massives", fait référence aux données trop importantes pour être lues ou interprétées individuellement par un chercheur, que ce soit des millions de tweets pour une élection, l'ensemble des manifestes politiques en Europe ou la littérature académique dans sa globalité. Par "science des données", ce sont l'ensemble des techniques de traitement de l'information qui permettent de structurer ces données massives et de les utiliser. Cette section vise à expliciter les différentes approches utilisées à travers la thèse doctorale. Par rapport aux "processus démocratiques", sont considérées les élections et la communication politique de manière générale, mais les résultats de la thèse peuvent aussi être appliqués à la consolidation de sondages et à la caractérisation du populisme. Le marqueur temporel de "l'ère d'Internet" souligne les changements structurels importants survenus au cours de la dernière décennie, où les géants de l'économie informationnelle comme Google, Facebook ou Twitter n'ont qu'une vingtaine d'années d'existence, et où les habitudes de consommation et de production de l'information des individus ont évolué.

Cette question de recherche peut être dérivée en plusieurs sous questions et hypothèses, auxquelles chaque article de la thèse tentera de répondre.

- Comment sont perçus les partis politiques au Québec sur les médias sociaux ?
- Comment se transposent les dynamiques électorales sur les médias sociaux en contexte canadien ?
- Comment pallier aux lacunes institutionnelles et au manque d'informations pour effectuer le suivi d'une élection ?
- Comment utiliser la science des données pour caractériser la montée du populisme ?

Concernant la première sous question de recherche, une importante portion de la littérature se concentre sur l'utilisation des médias sociaux en politique et cherche à comprendre le caractère prédictif de Twitter. Ici, l'objectif est de tirer profit de cette nouvelle source de

données afin de caractériser en temps réel les perceptions des utilisateurs des médias sociaux. Nous tenterons donc de mesurer la résonance de propositions politiques auprès des usagers de Twitter.

Ainsi, deux hypothèses seront abordées dans le premier article de la thèse :

— Hypothèse 1a : il est possible de mesurer en temps réel la résonance de sujets politiques sur Twitter.
— Hypothèse 1b : ces sujets politiques peuvent être associés préférentiellement aux partis politiques au cours d'une période électorale.

La seconde question de recherche veut analyser le contexte canadien. Plusieurs articles de recherche ont été publiés depuis la démocratisation de Facebook ou de Twitter, mais les études restent toutefois peu nombreuses. Pendant la durée de la thèse, des élections fédérales canadiennes ont eu lieu (2015), offrant l'opportunité de récolter des données alors inédites. De plus, les sondages électoraux au début de la campagne électorale présentaient le Parti Libéral du Canada en troisième position, ne prédisant pas la large victoire qui sera obtenue 80 jours plus tard.

Ainsi, le second article de la thèse permettra de tester les hypothèses suivantes :

— Hypothèse 2a : les dynamiques électorales se télescopent à travers les messages publiés sur Twitter.
— Hypothèse 2b : la persistance d'enjeux peut être observée.

Qu'arrive-t-il lorsque le niveau d'information concernant une élection ne permet pas d'en suivre la campagne électorale ? Lors des élections de 2015 au Nigeria, les sondages réalisés sur le terrain ne furent pas aussi impartiaux et appliquèrent d'autres règles que les standards associés aux institutions canadiennes ou américaines par exemple. Dans ce cas-ci, était-il possible d'utiliser les informations publiées sur Internet et provenant du Nigeria pour recréer de l'information ?

De cette prémisse fut alors dérivée deux hypothèses :

— Hypothèse 3a : les médias sociaux permettent une compréhension plus fine des évènements lors d'une campagne électorale.
— Hypothèse 3b : certains modèles économétriques sont plus adaptés que d'autres pour traiter les données issues des médias sociaux lors d'une élection.

Ces trois questions de recherche font l'objet de trois chapitres (articles de recherche). Le but de ces articles n'est pas de prédire les élections, puisque des limitations méthodologiques ont été relevées lors de la revue de littérature. Les articles visent à structurer, analyser et comprendre l'information publiée sur Twitter dans le cadre de ces élections, mais n'ont pas

portée à anticiper les différents résultats électoraux.

Finalement, depuis les années 2010, la question du populisme est un enjeu majeur dans la vie démocratique en général, et européenne plus particulièrement. Afin de mesurer le pouls de la population européenne, plusieurs sondages sont réalisés à intervalles réguliers et regroupés sous l'Eurobarometer. Les différentes élections nationales ou européennes mettent en avant des formations pouvant être étiquetées comme populistes. Certaines de ces formations obtiennent d'importants résultats aux élections, permettant à des partis d'extrême droite maintenant d'occuper des postes gouvernementaux en Europe. Toutefois, qu'en est-il des discours provenant des formations elles-mêmes ? Comment quantifier cette notion de populisme ?

Deux nouvelles hypothèses s'ajoutent à cette analyse :

— Hypothèse 4a : les partis politiques extrêmes ajustent leurs discours institutionnel dans le cadre d'élection pour ressembler aux partis de gouvernement.
— Hypothèse 4b : des thématiques communes rassemblent les partis politiques populistes en Europe.

Le tableau 3.1 suivant présente l'organisation de la thèse par article. Des quatre articles scientifiques, deux sont publiés, et deux sont soumis à publication au moment de la rédaction de cette thèse doctorale.

Afin de tester les différentes hypothèses de recherche, cette thèse se base sur des données inédites et met en oeuvre une méthodologie appropriée pour structurer les données massives récoltées. Cette méthodologie repose sur la science des données, et plus particulièrement sur l'économétrie, l'analyse de réseaux et le traitement du langage naturel. Nous détaillerons ces aspects dans la prochaine partie.

Tableau 3.1 Résumé de l'organisation de la recherche

	Article 1	Article 2	Article 3	Article 4
Titre original	The Public's Perception of Political Parties During the 2014 Québec Election on Twitter	The 2015 Canadian Election on twitter: A Tidy Algorithmic Analysis	Nigeria's 2015 Presidential Election: A Spatial and Econometric Perspective based on a Framing Strategy	Text-as-Data Analysis of Populist Parties versus Government Parties: To Blend or not to Blend?
Question(s) de recherche	Comment évolue la perception des partis politiques au XXIè siècle?	Comment se transposent les dynamiques électorales sur les médias sociaux en contexte canadien?	Comment pallier aux lacunes institutionnelles et au manque d'informations pour le suivi d'une élection?	Comment utiliser la science de données pour caractériser la montée du populisme?
Concepts clefs	Analyse de conversations, communication politique, médias sociaux, campagne électorale, science de données en sciences sociales	Analyse de médias sociaux, élections, LDA, sciences de données en sciences sociales, Canada	Médias sociaux, élections, Nigéria, analyse de texte, économétrie	Populisme, Europe, indice de similarité, LDA, politiques comparées
Objectifs de recherche	Mesurer la résonnance de sujets politiques sur Twitter et identifier les sujets associés aux différents partis politiques	Expliquer le déroulement de la campagne fédérale canadienne et développer une méthodologie d'analyse textuelle par LDA	Recréer des mesures de suivi électoral et développer une méthodologie économétrique pour le traitement des données de médias sociaux	Mesurer le rapprochement entre les doctrines de partis politiques et mettre en évidence les thématiques associées à l'extrême droite en Europe
État de la publication	Publié dans le Canadian Journal of Communication, Vol. 43 (2), pp. 245-263	Publié dans Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence, pp. 910-915	Soumis à International Journal of Social Research Methodology	Soumis à Proceedings of the National Academy of Sciences of the United States of America

3.2 Terrains d'observation et données récoltées

Les périodes électorales constituent des terrains de prédilection afin d'observer les différentes dynamiques des acteurs politiques à travers le prisme des données massives. De plus, l'importante fréquence à laquelle des élections ont lieu permet d'observer l'évolution de ces dynamiques à travers le temps. Deux types de données sont utilisés au cours de cette thèse. Dans un premier temps, des données acquises en temps réel depuis Twitter (constituant les données des articles 1 à 3). Le quatrième article de recherche utilise des données textuelles obtenues à partir de la base de données Manifesto Project.

3.2.1 Collecte de données

De par sa politique d'accès aux données, Twitter apparaît comme une source de données à haute valeur pour les études électorales. Afin de se connecter aux bases de données de Twitter, plusieurs étapes sont préalablement nécessaires. Après avoir enregistré un compte utilisateur, une application doit être créée afin d'obtenir les identifiants requis et permettre le téléchargement des données.

La librairie RCurl [Lang et al., 2018] permet d'effectuer des requêtes selon les protocoles http/https pour favoriser le téléchargement en temps réel de données provenant de serveurs informatiques. ROAuth [Gentry et Lang, 2015] sert d'interface d'identification pour se connecter à des serveurs via API et finalement la librairie rjson [Couture-Beil, 2018] transforme les données reçues en format .json en objets R (et inversement).

De plus, il existe plusieurs façons d'accéder aux données publiées sur Twitter. Entre autres, plusieurs API sont disponibles. L'API REST permet de remonter dans l'historique des messages publiés, mais de fortes limitations techniques subsistent : seuls 3 200 messages sont gardés en mémoire lors d'une requête "post" évènement, et l'on ne peut effectuer des requêtes demandant des messages publiés une semaine auparavant. En raison de ces deux limitations, nous n'utiliserons pas cette interface de programme pour assembler des bases de données. La deuxième méthode pour recueillir l'information auprès de Twitter consisterait à faire appel à l'API Firehose, qui regroupe l'intégralité des données mises en ligne par la plateforme. Cette dernière option n'est pas envisageable, à cause du volume de messages publiés (500 millions quotidiennement) et du fort coût associé à l'utilisation de cette API. Un troisième protocole d'interface de programmation, Streaming API, est plus versatile puisqu'il permet aux programmeurs d'effectuer des requêtes en temps réel auprès de Twitter. Un champ de recherche est acheminé puis une série de messages sont par la suite renvoyés sous format .json. Plusieurs programmes développés en Python, Ruby et C++ ont été mis au point afin

d'aller recueillir ces informations et construire des bases de données appropriées. Toutefois, le langage de programmation R offre une série de librairies permettant la connexion aux bases de données de Twitter, notamment twitteR [Gentry, 2015], streamR [Barbera, 2018] et rtweet [Kearney, 2018b]. Cette méthodologie avait été développée lors d'un précédent travail académique [Sanger, 2014].

Pour le premier article de la thèse, les données issues de Twitter furent collectées entre le 17 mars 2014 et le 7 avril 2014. À partir des mots clics #assnat (Assemblée Nationale), #polqc (politique québécoise) et #qc2014 (Québec 2014), un total de 672 497 messages furent considérés. En utilisant les plateformes des différents partis politiques, 31 thématiques de campagne ont pu être identifiées, représentant 157 916 messages. De plus, la campagne électorale fut structurée par deux débats télévisés, permettant d'observer les dynamiques électorales durant trois périodes différentes.

Le deuxième article a permis la constitution d'une base de données de 3 498 633 tweets provenant de 218 255 utilisateurs uniques. À nouveau, les mots clics utilisés durant la campagne électorale furent sélectionnés, notamment #exln42 (42è élection fédérale), #cdnpoli (Canadian politics) et #polcan (politique canadienne). En moyenne, le nombre de messages collectés par jour fut de 38 000 entre le 2 août 2015 et le 18 octobre 2015, mais atteignit de plus importantes proportions lors des débats télévisés et le jour du scrutin électoral.

Les élections nigérianes de 2015 ont servi de terrain d'étude pour le troisième article de la thèse. Au cours de cette élection, deux candidats principaux s'affrontèrent, le président sortant Goodluck Jonathan et le challenger Muhammadu Buhari. Durant les 27 derniers jours des élections, deux bases de données furent constituées. La première avait pour critère de sélection la géolocalisation des messages publiés sur Twitter. Tous les messages émis depuis le Nigeria furent donc collectés et constituèrent un ensemble de 555 811 messages. La seconde base de données reprend la méthodologie de cadrage précédemment utilisée et concerne plus de 3,8 millions de messages mentionnant le Nigeria, Boko Haram, Buhari ou Goodluck Jonathan par exemple. Cette dernière base de données fut réduite à 1,541 millions de messages en ne prenant en compte que les messages mentionnant les chefs de partis politiques et les mots clics des élections, soit #nigeria2015 et #nigeriadecides.

Finalement, le quatrième article utilise les textes des plateformes politiques comme source première de données. À travers le Manifesto Project [Lehmann *et al.*, 2018], accessible via une librairie en R - ManifestoR [Lewandowski *et al.*, 2018], les textes utilisés proviennent des 28 pays de l'Union européenne (27 si l'on ne compte pas Malte, pays absent de la base de données) entre 2000 et 2018. Des méthodes de structuration des données sont utilisées, puisque l'ensemble des textes accessibles correspondent à un total de plus de 12 millions de

mots.

3.2.2 Structuration de données

Une étape importante du travail de structuration de données est de considérer les données dans une forme "rangée" (en anglais *tidy*). Ce formatage permet le traitement rapide de tous types de données, qu'elles soient textuelles, numériques, catégoriques ou manquantes. Pour être plus précis, chaque base de données doit être rangée selon la description proposée par [Wickham, 2014] : chaque variable est associée à une colonne de la base de données, chaque observation est une ligne, et chaque cellule correspond à une valeur.

Cette structuration de données offre une grande versatilité d'approche. En effet, des textes entiers peuvent être analysés de cette manière-là, notamment à partir des librairies développées telle que tidytext [Queiroz *et al.*, 2018,Silge et Robinson, 2017,Silge et Robinson, 2016]. Les manipulations de données subséquentes deviennent alors des processus standards et réplicables. De plus, l'interopérabilité des librairies au sein du tidyverse [Wickham et RStudio, 2017b] fait en sorte que les données "rangées" soient le format de base d'utilisation de ces méthodes de traitement.

3.2.3 Manipulation de données

Une composante importante des données récoltées est la dimension temporelle leur étant associée. Par exemple, lorsqu'un tweet est téléchargé, ce sont plusieurs dizaines de types d'informations qui sont aussi collectés à travers ses métadata. Les données liées à la date et aux heures sont des données dont la manipulation peut s'avérer complexe, nécessitant ainsi la librairie lubridate [Spinu *et al.*, 2018] qui facilite ce genre de manipulations.

Pour modifier les bases de données, plusieurs possibilités sont envisagées. D'abord, la librairie reshape2 [Wickham, 2017] aide à l'agrégation et au développement de bases de données en fonction des différentes variables présentes. Suite à cette librairie a été développée tidyr [Wickham *et al.*, 2018c] qui offre une évolution des commandes précédentes en incorporant l'utilisation des symboles de transferts de fonctions (%>%). dplyr [Wickham *et al.*, 2018b] est une série de fonctions permettant la manipulation de données à partir de langage quasi naturel, rendant le code informatique développé facilement interprétable pour l'observateur externe. Finalement, l'ensemble de ces différentes fonctions sont regroupées au sein d'une même librairie principale, tidyverse [Wickham et RStudio, 2017b]. À partir d'un esprit commun, ce sont plusieurs librairies de traitement de données en R qui sont basées sur les mêmes principes de données rangées afin de permettre l'interopérabilité entre ces librairies.

3.3 Science des données

La méthodologie développée à travers cette thèse repose sur les techniques de science des données. À travers le langage de programmation R et l'utilisation de l'environnement de développement proposé par RStudio, les processus reproductibles de recherche sont appliqués, favorisant ainsi le transfert de données, la transparence des méthodologies utilisées et assurant la réplicabilité des résultats. Le caractère réplicable de la recherche en sciences sociales est un défi auquel la science des données permet de répondre [King, 2011].

Cette section détaille donc les techniques employées, notamment les différents modèles économétriques, l'analyse de réseaux, le traitement du langage naturel et la visualisation des données, les algorithmes structurant les méthodes d'analyses et les librairies utilisées en R.

3.3.1 Modèles économétriques utilisés

Un des constats de la littérature concerne le manque de méthodes structurantes et robustes pour traiter les données massives. Les approches promues varient selon les auteurs et les groupes de recherche. Outre une contribution thématique concernant l'analyse des différentes élections survenant depuis le début de la thèse, une contribution méthodologique est envisagée par le développement de méthodes robustes pour analyser économétriquement les données publiées sur Twitter.

Les modèles logistiques constituent des modèles adéquats pour l'interprétation de l'effet de différentes variables en contexte électoral. En effet, par la nature du choix proposé aux électeurs lors d'un vote, la variable explicative peut être une variable binaire (choix entre deux candidats) ou une variable catégorique (entre plusieurs alternatives). Dans le premier cas, on envisagera un modèle logistique binaire, tandis que pour le second ce sera un modèle logistique multinomial.

Ici, les modèles logistiques seront utilisés afin d'appréhender l'information non structurée issue de ces données massives. En premier lieu, les données non structurées doivent être codées afin d'en retirer l'information qualitative contenue dans un message. En l'occurrence, nous chercherons à identifier un candidat, un parti ou une thématique mentionnée à travers le texte d'un tweet. Dans le cas des thématiques électorales, nous pouvons supposer qu'elles soient mutuellement exclusives, puisqu'en 140 caractères (ou 280 depuis les changements de politique d'utilisation de Twitter en novembre 2018) un utilisateur ne s'exprimera que sur une thématique préalablement identifiée.

Concernant les modèles économétriques à utiliser, dans le cas où la variable expliquée est une variable sous la forme binaire (un parti par rapport à un autre, une thématique par

rapport à une autre), il importe de choisir entre un modèle probit ou logit. Comme démontré par la littérature, l'interprétation des modèles probit peut prêter à confusion à cause de la fonction de transfert utilisée et de la nature des termes d'erreur [Dow et Endersby, 2004]. Ainsi, un modèle logit binaire permettrait de caractériser l'influence de variables dans ce cas de figure-ci.

Ensuite, si la variable expliquée prend la forme d'une variable multinomiale, il est important de préciser si cette variable possède ou non un ordre de classification (plusieurs choix différents qui ne sont pas ordonnés les uns par rapport aux autres ou un classement ordonné entre ces différents choix). Si l'on désire expliquer l'importance de thématiques électorales au sein des conversations sur Twitter, un ordre préférentiel est induit par la classification de ces thématiques par les utilisateurs de la plateforme en ligne. Un proxy de classification pourrait être l'importance de la thématique en nombre de messages publiés quotidiennement. Cette classification par les utilisateurs procure un avantage dans la compréhension des dynamiques électorales, puisqu'elle n'est pas influencée par des biais méthodologiques attribués par les sondages.

Ainsi, si la variable expliquée peut être rangée par ordre, il faudra employer un modèle logit multinomial ordonné. Si la variable expliquée n'est pas rangée préférentiellement, il faudra alors employer un modèle logit multinomial. Des modèles multinomiaux mixtes seraient aussi envisageables afin de contrôler la distribution des termes d'erreur [Glasgow, 2001, Jaeger, 2008].

Afin d'apporter une profondeur d'analyse plus poussée, des caractéristiques des utilisateurs peuvent être mises en évidence. [Barberá *et al.*, 2015] utilise l'idéologie partisane comme variable d'analyse supplémentaire. Ce type de variable peut donc être extraite des comptes utilisateurs de Twitter en fonction de leurs listes d'abonnements. Outre cette variable, des données reliées au nombre d'abonnés et des variables binaires peuvent être utilisées, mesurant par exemple le jour de publication, la zone géographique, le type d'expertise des utilisateurs... Finalement, ces variables supplémentaires peuvent être utilisées en interaction, et l'estimation de l'effet de ces termes d'interaction devra prendre en compte les effets partiels des variables combinées (sans oublier l'effet des potentiels termes d'interaction de l'équation des modèles économétriques) [Ai et Norton, 2003, Greene, 2010, Powers, 2005].

Il est intéressant de noter que la profondeur et la granularité d'analyse des données massives permet de mettre en évidence des dynamiques électorales ayant une incrémentation temporelle d'une journée (appui à un candidat, thématique importante) mais allant aussi à la minute près (réaction face à une proposition politique lors d'un débat télévisé par exemple).

3.3.2 Analyse de réseaux

Concernant les données issues de Twitter, lorsqu'un individu retweet un autre individu, c'est à dire retransmet à sa liste de contacts un message écrit par une personne tierce, un lien se crée entre les deux comptes utilisateurs. C'est une des interactions centrales du réseau social Twitter. En cartographiant l'ensemble des liens entre utilisateurs, il est alors possible de visualiser les réseaux de transmission d'informations sur les médias sociaux. Des techniques similaires ont été utilisées pour mesurer la crédibilité d'individus publiant des messages à caractère financier sur Twitter [Marcellis-Warin *et al.*, 2017] ou pour caractériser les liens entre les conseils d'administration de compagnies financières à travers le monde [Warin et Sanger, 2018].

Plusieurs librairies en R sont utilisées pour analyser ce genre de données de réseaux. igraph [Csardi et Nepusz, 2006, Csardi et al., 2018] permet de gérer les données associées à de simples réseaux jusqu'à des réseaux les plus complexes. Cette librairie extrêmement exhaustive procure l'ensemble des informations analytiques nécessaires à l'analyse de réseaux. network [Butts *et al.*, 2018] permet la conversion et le traitement de données associées aux réseaux. Plusieurs librairies offrent des algorithmes similaires de manière intégrée, notamment statnet [Handcock *et al.*, 2008, Handcock *et al.*, 2016] ou sna [Butts, 2016]. La librairie ape [Paradis *et al.*, 2004, Paradis *et al.*, 2018] possède les formules nécessaires à la mesure de communauté au sein de réseaux.

Pour être plus précis et en reprenant les propriétés mathématiques expliquées dans l'article de [Warin et Sanger, 2018], plusieurs mesures permettent de quantifier les caractéristiques d'un noeud au sein d'un réseau. La plupart des mesures de réseaux développées dans ces librairies reprennent les formules qui sont ici explicitées.

Ainsi, la centralité de degré équivaut au nombre de connections établies par rapport à un point dans un réseau [Freeman, 1977, Nieminen, 1974]. Variant entre [0 ;1], la centralité de degré d'un noeud p_k est défini par :

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \tag{3.1}$$

où $a(p_i, p_k) = 1$ si p_i et p_k sont reliés entre eux, et 0 dans le cas contraire.

La centralité de proximité permet de mesurer le rapprochement entre deux points au sein d'un réseau. Plus un point sera central par rapport aux autres points de ce réseau, plus il sera proche physiquement de ces points. Introduit par [Sabidussi, 1966], la centralité de proximité s'évalue telle que :

$$C_P(p_k)^{-1} = \sum_{i=1}^{n} d(p_i, p_k) \tag{3.2}$$

où $d(p_i, p_k)$ est la distance entre deux noeuds du réseau.

La centralité d'intermédiarité mesure la probabilité d'un noeud à se trouver sur le plus court chemin entre deux autres noeuds [Freeman, 1977, Kolaczyk et Csárdi, 2014] telle que :

$$C_I(p_k) = \sum_{p_k \neq p_j \neq p_i \in V} \frac{\sigma(p_i, p_j|p_k)}{\sigma(p_i, p_j)} \tag{3.3}$$

avec $\sigma(p_i, p_j|p_k)$ le nombre total de chemins reliant deux noeuds p_i et p_k où p_k se trouve et $\sigma(p_i, p_j)$ le nombre total de chemins les plus courts entre p_i et p_j.

Finalement, la centralité de valeurs propres (eigenvalue centrality) peut être mesurée [Bonacich, 1972] telle que :

$$C_{Ei}(p_k) = \alpha \sum_{\{p_k, p_i\} \in V} C_{Ei}(p_i) \tag{3.4}$$

où $C_{Ei} = (C_{Ei}(1), ..., C_{Ei}(N))^T$ est la solution au problème de valeurs propres $AC_{Ei} = \alpha^{-1} C_{Ei}$ et où A est la matrice adjacente modélisant le réseau [Kolaczyk et Csárdi, 2014].

Les représentations graphiques peuvent utiliser les algorithmes de [Kamada et Kawai, 1989] ou de [Fruchterman et Reingold, 1991] selon les visualisations désirées afin de mettre en évidence les structures des différents réseaux.

3.3.3 Visualisation

Une autre force du langage de programmation R est la diversité des librairies permettant la visualisation de données. Ainsi, sous l'égide d'une librairie centrale, ggplot2 [Wickham, 2010, Wickham, 2016, Wickham *et al.*, 2018a], les visualisations les plus simples comme les plus complexes sont possibles. Cette librairie s'intègre parfaitement à un système de données en format rangé, facilitant et raccourcissant le temps de traitement d'un jeu de données.

D'autres librairies sont utilisées en complément, notamment pour fournir des couches graphiques aux graphiques produits, tels que ggthemes [Arnold *et al.*, 2018], ggsci [Xiao et Li, 2018], scales [Wickham et RStudio, 2017a] ou wordcloud [Fellows, 2014].

3.3.4 Traitement du langage naturel

Les dernières techniques considérées dans cette thèse doctorale sont des techniques d'analyses textuelles. En fonction des données utilisées, plusieurs stratégies peuvent être employées. En effet, les données varient entre de simples tweets à des textes de plusieurs pages, et donc nécessitent un traitement spécifique.

Les expressions régulières, utilisées avec les librairies stringr [Wickham et RStudio, 2018] et tm [Feinerer *et al.*, 2018] permettent de rechercher au sein d'un texte des chaînes de caractères particulières. Ainsi, isoler un ensemble de tweets mentionnant un candidat en particulier s'avère une tâche routinière à l'aide de ces librairies.

Des environnements plus développées existent, notamment RTextTools [Collingwood *et al.*, 2013, Jurka *et al.*, 2014] qui offre une série d'outils d'analyse textuelles intégrés. tidytext [Queiroz *et al.*, 2018,Silge et Robinson, 2017,Silge et Robinson, 2016] applique la méthodologie de données rangées aux textes, facilitant l'utilisation de fonctions issues des librairies du tidyverse.

Quatre approches sont notamment développées dans cette thèse doctorale. Dans un premier temps, une analyse de sentiment peut être effectuée sur les données textuelles. Chaque texte est associé à une émotion en particulier parmi la colère, le dégoût, la peur, la joie, la tristesse ou la surprise. Un lexique composé de 1 542 mots sert de référence avec respectivement 355, 70, 195, 553, 274 et 95 mots associés à chaque émotion.

Afin de refléter un score de sentiment à chaque texte, les données textuelles sont d'abord nettoyées de mots vides de sens puis comparées aux différents lexiques. En suivant un algorithme bayésien, des scores de sentiment sont donc obtenus tels que :

$$\left| ln(\frac{nb_{emotion}}{m_{emotion,lexique}}) \right| \tag{3.5}$$

La seconde approche est de considérer la polarité des textes, c'est à dire d'établir un indice de mesure sur une échelle pouvant être négative ou positive. Plusieurs stratégies peuvent être utilisées en fonction des dictionnaires de références.

Une première base lexicale de [Hu et Liu, 2004] regroupe 6 788 mots classés positivement ou négativement de manière binaire. Un score de polarité est par la suite associé à chaque élément textuel tel que :

$$polarite_{Hu.et.Liu;i;j} = positif_{i;j} - negatif_{i;j} \tag{3.6}$$

Un second lexique peut être utilisé, faisant lui appel à 2 476 mots de référence [Nielsen, 2011]. La particularité de ce lexique de référence est que les mots ne sont pas classés de manière binaire mais plutôt sur une échelle allant de -5 (très négatif) à +5 (très positif). Cela permet d'obtenir des mesures plus fines de polarité, notamment à travers l'occurrence des différents mots référencés, tels que :

$$polarite_{Nielsen;j} = ScorePolarite_{j;k} * NombreOccurences_{j;k} \tag{3.7}$$

Une troisième approche consiste à réaliser une analyse de sujets au sein des bases de données. Cette analyse par allocation de Dirichlet latente (LDA, pour Latent Dirichlet Allocation) permet de regrouper les différents éléments d'un texte selon des catégories similaires. En conjonction avec la librairie topicmodels [Grün et Hornik, 2017, Hornik et Grün, 2011], il est donc possible de déterminer un nombre idéal de sujets composant un texte, puis de regrouper les éléments de ce texte selon ces différentes catégories.

L'avantage d'une telle méthode est que le biais de sélection du chercheur est réduit puisque cette méthodologie est essentiellement algorithmique et non sujet au jugement d'un observateur externe. Une optimisation du nombre de sujette différents à spécifier peut aussi être effectuée, continuant à réduire l'implication du chercheur [Asuncion *et al.*, 2009].

Finalement, la dernière méthodologie d'analyse textuelle en science des données est une analyse de similarité entre textes. En effet, afin de mesurer et de produire des indices de similarité de programmes politiques en Europe, il est nécessaire de construire des indices comparables entre les pays malgré l'emploi de langues différentes dans les bases de données.

Pour cela, l'indice de Jaccard est employé. Initialement utilisé en botanique au début des années 1900 [Jaccard, 1902b], cet indice compare deux ensembles de données et indique le pourcentage d'éléments similaires parmi ces deux ensembles tel que :

$$J(ensemble_1, ensemble_2) = \frac{|ensemble_1 \cup ensemble_2|}{|ensemble_1 \cap ensemble_2|} \tag{3.8}$$

Cette mesure a été utilisée dans plusieurs domaines, notamment en affaires internationales pour mesurer la persistance de traités internationaux [Alschner *et al.*, 2017].

Cette technique possède aussi deux avantages. Le premier est que c'est une technique neutre au langage, c'est à dire qu'elle ne nécessite pas de lexique de comparaison (comme les analyses de sentiment par exemple) et qu'elle peut être utilisée peu importe la langue étudiée. Ainsi, des études en arabe [Al-Kabi et Al-Sinjilawi, 2007, Thabtah, 2008] et en thaïlandais [Niwattanakul *et al.*, 2013] ont été réalisées utilisant les indices de Jaccard. L'autre avantage est que

cette technique peut prendre en compte le contexte lexical des données. Une des critiques des approches par paquets-de-mots (bag-of-words) est que le sens complet d'une phrase n'est pas nécessairement pris en compte (notamment dans le cadre de mesures de polarité). En suivant la technique de [Alschner et Skougarevskiy, 2016] qui consiste à diviser un texte en groupes de cinq caractères (unigram 5-characters), il est alors possible d'interpréter le contexte lexical des textes.

Finalement, la librairie textreuse [Mullen, 2016] est utilisée pour fournir l'indice de similarité entre deux textes considérés.

3.4 Organisation de la recherche

Les quatre articles constituant la thèse ont été publiés ou sont soumis à publication. Le choix des revues a été influencé par les thématiques étudiées ou les méthodologies utilisées. Ainsi, l'article 1 intitué **Public's Perception of Political Parties During the 2014 Québec Election on Twitter** a été publié en 2018 dans le Canadian Journal of Communications. La venue de prochaines élections provinciales (automne 2018) et l'angle d'analyse du domaine de la communication justifièrent le choix de ce journal.

Le second article intitulé **The 2015 Canadian Election on Twitter : A Tidy Algorithmic Analysis** a été présenté lors de la 4th International Conference on Computational Science & Computational Intelligence (CSCI) de Las Vegas aux États-Unis. Ce travail a fait l'objet d'un papier de recherche publié au printemps 2018 par IEEE Xplore. IEEE s'adresse à un public généralement issu du domaine informatique. Toutefois, les techniques mises en valeur dans ce papier de recherche promeuvent l'utilisation de sciences de données appliquées aux sciences sociales.

Le troisième article de la thèse, intitulé **Nigeria's 2015 Presidential Election : A Spatial and Econometric Perspective based on a Framing Strategy** a quant à lui été soumis pour publication à l'hiver 2019 auprès du journal International Journal of Social Research Methodology. La forte coloration méthodologique du troisième article de la thèse fit de ce journal un choix approprié.

Quant au quatrième article de recherche intitulé **Text-as-Data Analysis of Populist Parties versus Government Parties : to Blend or not to Blend ?**, il fut soumis à publication à l'hiver 2019 auprès du journal Proceedings of the National Academy of Sciences of the United States of America.

Le tableau 3.2 rappelle les différents éléments constituant le cadre méthodologique de chaque article de la thèse.

Tableau 3.2 Synthèse des méthodologies utilisées par article

	Article 1	Article 2	Article 3	Article 4
Objectif principal de l'article	Mesurer la résonnance de sujets politiques sur Twitter et identifier les sujets associés aux différents partis politiques	Expliquer le déroulement de la campagne fédérale canadienne et développer une méthodologie d'analyse textuelle par LDA	Recréer des mesures de suivi électoral et développer une méthodologie économétrique pour le traitement des données de médias sociaux	Mesurer le rapprochement entre les doctrines de partis politiques et les thématiques associées à l'extrême droite en Europe
Méthodologies	Économétrie	Analyse de sentiment, LDA	Cartographie, Économétrie	Analyse de similarité, LDA
Unité(s) d'analyse	Élections 2014 au Québec et Twitter	Élections 2015 au Canada et Twitter	Élections 2015 au Nigéria et Twitter	Élections parlementaires en Europe (2000-2017)
Données	672 497 messages publiés sur Twitter sous les mots-clics #assnat, #polqc, #qc2014 (8 mars 2014 - 7 avril 2014)	3 498 633 messages publiés sur Twitter sous les mots-clics #exln42, #cdnpoli, #polcan (2 août 2015 - 18 octobre 2015)	555 811 messages géolocalisés publiés sur Twitter provenant du Nigéria et 1 541 000 messages sous les mots-clics #nigeria2015, #nigeriadecides (1 mars 2015 - 27 mars 2015)	676 manifestes politiques provenant de 27 pays en Europe (2000 - 2018), soit l'équivalent de 12 041 159 mots
Sources des données	API de Twitter	API de Twitter	API de Twitter	Manifesto Project

CHAPITRE 4 ARTICLE 1: PUBLIC'S PERCEPTION OF POLITICAL PARTIES DURING THE 2014 QUEBEC ELECTION ON TWITTER

4.1 Présentation de l'article

Référence. Sanger, W. et Warin, T. 2018. Public's Perception of Political Parties During the 2014 Quebec Election on Twitter. Canadian Journal of Communication. Vol. 43 (2). pp. 245-263.

Cet article est coécrit avec Thierry Warin. Il a été accepté pour publication en septembre 2017 dans la revue Canadian Journal of Communication. Il a été publié au printemps 2018.

Ce premier article de la thèse concerne l'élection québécoise de 2014. Le Parti Québécois, alors au pouvoir dans un gouvernement minoritaire dirigé par Pauline Marois, décide de déclencher une campagne électorale. Un mois plus tard, ce fut le Parti Libéral, dirigé par Philippe Couillard, qui remporta l'élection et put former un gouvernement majoritaire.

Entre l'élection de 2014 et la précédente (2012), plusieurs caractéristiques diffèrent. Par exemple, le nombre de débats est différent et le gouvernement au pouvoir en 2012 est majoritaire alors que le gouvernement au pouvoir en 2014 est minoritaire. Ces considérations permettent de faire émerger plusieurs hypothèses en termes de stratégies électorales. Quant à la perception des individus sur les médias sociaux, change-t-elle en fonction des propositions électorales, des débats, des événements marquant la campagne ? C'est dans ce contexte que se situe cet article de recherche.

L'objectif principal de l'article est de comprendre la manière dont sont perçus les partis politiques au Québec sur les médias sociaux. Pour cela, cet article associe préférentiellement les thématiques de campagnes aux différents partis politiques à travers l'analyse de 670 000 messages publiés sur Twitter. Des modèles économétriques ont été utilisés pour associer les thèmes liés à l'indépendance du Québec, l'économie, la société et l'environnement aux quatre partis politiques en présence. Le résultat de l'article a été de développer un cadre méthodologique afin de structurer l'information issue des médias sociaux lors d'une élection.

4.2 Abstract

Background. This article investigates how to extract signals from social media (Twitter) concerning political parties during an election.

Analysis. 670,000 messages were collected during the 2014 Québec election regarding each political party using a framing strategy. After associating each message to one of the four main topics of the campaign, two logistic models were developed to describe the election. While having been set by the incumbent party, the topic of "Independence" was not the most important topic of the campaign ("Economy" and "Society" were). When dominating in terms of mentions, each party was associated to a topic, and such association changed during the campaign.

Conclusion and implications. From a practical standpoint, the findings of this article could be used to implement a framework to understand political campaigns dynamics through social media.

Keywords. Conversation analysis ; Political communication ; Social media ; Electoral campaign ; Social Data Science

4.3 Résumé

Contexte. Cette recherche est axée sur la manière de structurer les signaux issus des médias sociaux (Twitter) en contexte politique.

Analyse. Nous avons collecté 670 000 messages concernant l'élection québécoise de 2014 en utilisant une stratégie de cadrage. Chaque message fut associé à une thématique de campagne, puis deux modèles logistiques furent utilisés pour décrire les élections. Ainsi, alors que le thème de l'indépendance fut mis à l'avant par le parti sortant, ce sont les messages reliés à l'économie et à la société qui furent les plus importants. Chaque parti fut associé préférentiellement à une thématique lorsqu'il domina en termes de mentions, et nous observons une évolution de cette association au cours de la campagne électorale.

Conclusions et implications. Les résultats de cette recherche peuvent servir de cadre analytique pour structurer l'utilisation de données massives en contexte électoral.

Mots clés. Analyse de conversations ; Communication politique ; Média sociaux ; Campagne électorale ; Science des données en sciences sociales

4.4 Introduction and the context of the 2014 Québec election

March 5, 2014, marked the start of a new general election in Québec. After 18 months in power with a minority government, the leader of Parti Québécois, Pauline Marois, sought to be elected with a majority government. A month later on April 7, the leader of the Québec Liberal Party (QLP), Philippe Couillard, won the race with a majority of the votes, allowing him to form a solid government for the next four years. It is interesting to note that Pauline Marois did not have to trigger a general election ; there was no particular pressure to do so. Therefore, if she decided to run a campaign, it was because the polls were favourable to her political party. Moreover, as the incumbent, Pauline Marois thought, rightfully, that she would have an advantage over the other parties in terms of setting up the campaign agenda. Indeed, the incumbent is the only party that knows with certainty if and when citizens will be called to vote and when the campaign will start. It provides at least two competitive advantages : on the one hand, the incumbent can put forward some political or societal topics even before the other parties know that there will be an election in the near future, and on the other hand, the incumbent can consider its own stance in the polls and decide the best moment for its own re-election.

In short, based on these two competitive advantages, Pauline Marois was not considered as taking too big a risk, and rationally thought that the outcome would be one of only two options : 1) she would win with a majority government, or 2) she would win, but stay as a minority party leading the government. The latter option corresponds obviously to the worst-case scenario. Therefore, the only risk she was taking was to stay in the same situation or to improve it. In this context, the decision was easy to make : Québec would have a general election in April, 2014, after a short month of campaigning.

However, the results did not turn out as expected for Pauline Marois. Indeed, Philippe Couillard's party not only won the election and was put in the position of forming a government, but more importantly it was a majority government. The question is thus : what are the reasons that led to such a tremendous and unexpected reversal of fate ? Political scientists will study this election with particular interest. The aim of this article is to see whether social media can be used to study elections in the Québec context. The goal is not to predict the results of an election, but to try to extract the trends from the conversations on social networks and have a better explanation of what happens during a campaign. This article focuses only on Twitter. The reasons are essentially twofold : 1) the data are more easily available compared to other platforms, and 2) it is possible to compare this 2014 election to the 2012 election.

Therefore, the research question is : can the conversations that took place on Twitter during the electoral campaign help understand the unexpected reversal of fate for Pauline Marois' party ?

This research question could be a subset in several sub-questions to analyze the conduct of the electoral campaign :

— In this day and age of social networks in which the pace of conversations is dramatically increased, 1) does it matter for a party to have a program in order to try to set the campaign agenda, or 2) is it better to be prepared with the buzzwords that will emanate from the conversations on online platforms ?
— Is there a moment or a period during a campaign that matters more ?
— If the incumbent has a competitive advantage when starting the campaign, how long does it keep it ?
— Is there an optimal number of debates to have during a one-month campaign ?
— Do debates help change the pace and/or the themes of a campaign ?

It is already interesting to note that to answer these questions in a traditional manner, e.g., through polls, would be in fact very expensive. By using social networks—and Twitter in particular—it should be possible to address these questions in a very inexpensive way : indeed, the platform allows any user to connect to its Application Programming Interface (API) in order to access tweets publicly published, which helps build databases dedicated to answer those questions.

Many events occurred during the Québec electoral campaign. Some were expected and some differed from the previous campaign in 2012. Among the expected circumstances, there were four main opposing parties in action : the Parti Québécois (PQ, leader : Pauline Marois), the Québec Liberal Party (QLP—PLQ in French and in the rest of the article, leader : Philippe Couillard), the Coalition Avenir Québec (CAQ, leader : François Legault) and Québec Solidaire (QS, leaders : Françoise David and Andrés Fontecilla).

Among the new events were two debates on March 18, 2014, and March 27, 2014. Those debates confronted each leader on different topics, such as Québec's independence, ethical behaviours in the government, economics, and society. The two-debates approach was a major difference from the previous campaign held in 2012, which had only one debate. Another important difference is the adoption of Twitter by the population as a tool for commenting and sharing information or personal thoughts online and in real time. Finally, former CEO and President of Quebecor Pierre Karl Péladeau announced his candidacy in Saint-Jérôme's electoral district for the Parti Québécois 29 days before the election day. Known for his favourable opinion on Québec's independence, this unexpected announcement from Pauline

Marois set the table for discussions on this topic at the beginning of the electoral campaign.

To start, this article compares both campaigns (2012 and 2014) in order to illustrate the changes in the use of Twitter. Between both electoral campaigns, the number of messages published on Québec politics increased, and peaked at the time of television debates. Figure 4.1 illustrates the number of messages sent on a daily basis for both electoral campaigns. In 2012, the hashtags considered were #assnat, #polqc, and #qc2012, and in 2014 they were #assnat, #polqc, and #qc2014.

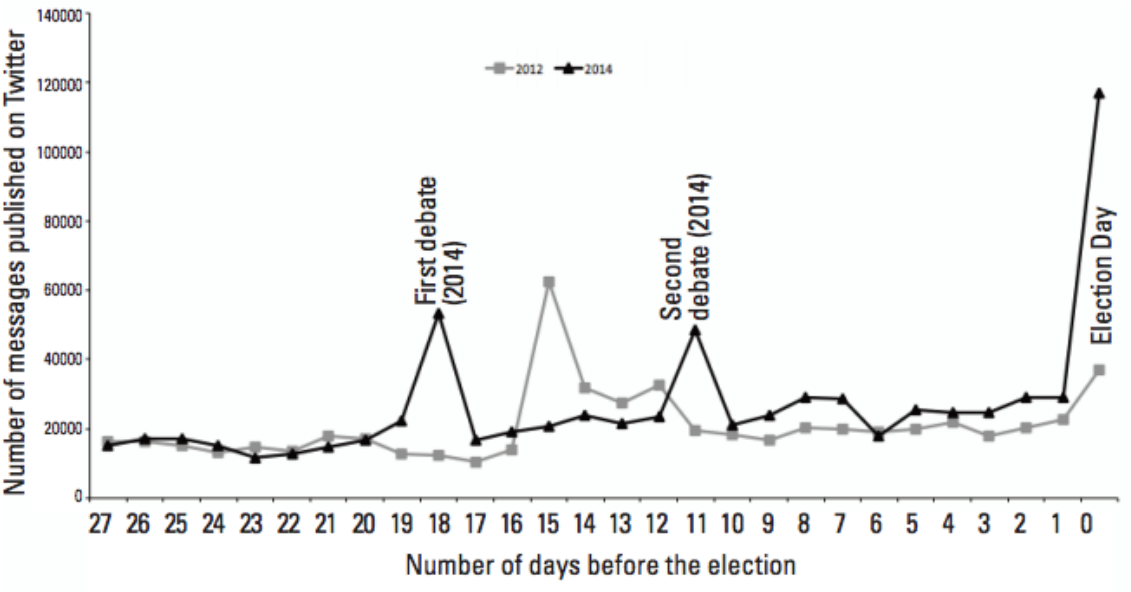

Figure 4.1 Number of messages published on Twitter

Electoral campaigns are monitored by polls published during this particular period of time. However, polls are only available a few days apart from each other and with some significant delays, which is not the case with a social network media such as Twitter. Indeed, social media allow access to their data through different APIs, offering the opportunity to analyze in near realtime the reactions of the population. More precisely, a poll realized by the polling institute CROP from March 12 to March 16 and released in the press on March 18, predicted that the Québec Liberal Party was suddenly running ahead in terms of vote intentions compared to the Parti Québécois. This change in lead was observed three days ahead, on March 15. In fact, by observing the volume of tweets mentioning both party leaders, Philippe Couillard was outperforming Pauline Marois. This advantage in terms of presence on Twitter persisted until the end of the electoral campaign. This last case is an example of how to assess a finer level of information through unstructured data (tweets) in a faster pace than traditional polls.

Figure 4.2 plots the share of messages mentioning political party's leaders in 2012. Françoise David and Amir Khadir (co-leaders of QS) have a very low representation on Twitter in 2012. Comparing Figure 4.2 with Figure 4.3, it is apparent that Françoise David improved her presence in 2014. In 2012, Jean Charest started high and then his share dropped dramatically. On the contrary, Pauline Marois and François Legault saw an increase in their share in the last period of the electoral campaign.

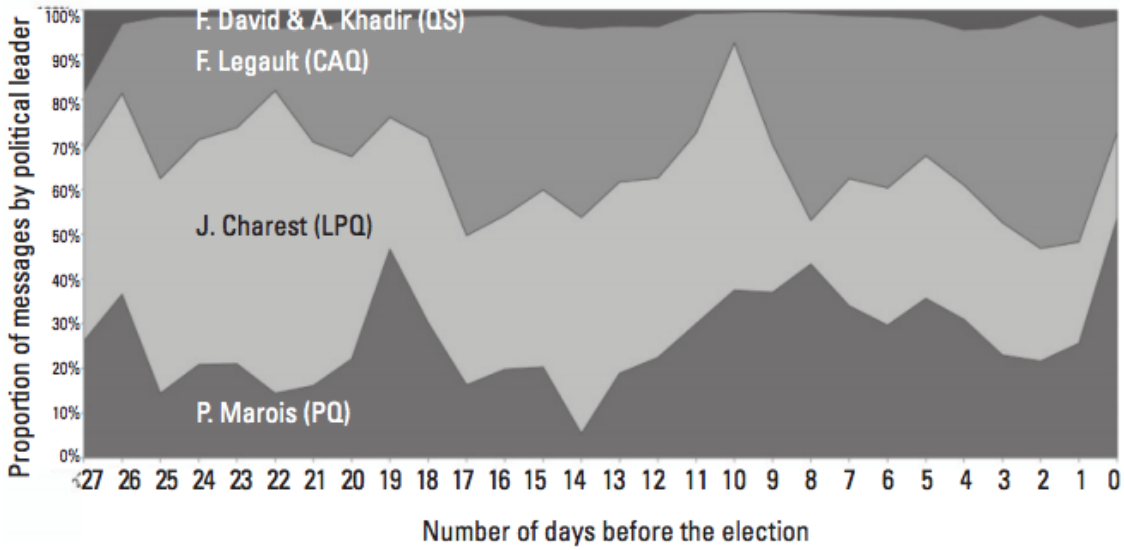

Figure 4.2 Proportion of messages by political leader, 2012

During the 2014 electoral campaign (see Figure 4.3), Philippe Couillard and Pauline Marois started high in terms of presence on Twitter. It is also interesting to note that François Legault stayed at a low level for a very long time. For him, the change happened after the second debate when the economy started to resonate more with the electorate on Twitter compared to the previous period.

The research question is : can the conversations that took place on Twitter during the electoral campaign help us understand the unexpected reversal of fate for Pauline Marois' party ? In order to do so, the evolution of topics during the electoral campaign on Twitter is examined. The focus is on establishing which party is associated with which particular topic, and if that stays stable until the election day. More than 670,000 messages were collected during the last 28 days of the electoral campaign, and a text analysis was performed on the content of these messages. This article presents a review of the literature to develop a framework to analyze this research question. Methodology and results are then presented, particularly the explicit methodology used to extract messages and to build the dataset. The econometric approach developed in this article is explained and, finally, it comments on the use of Twitter as a

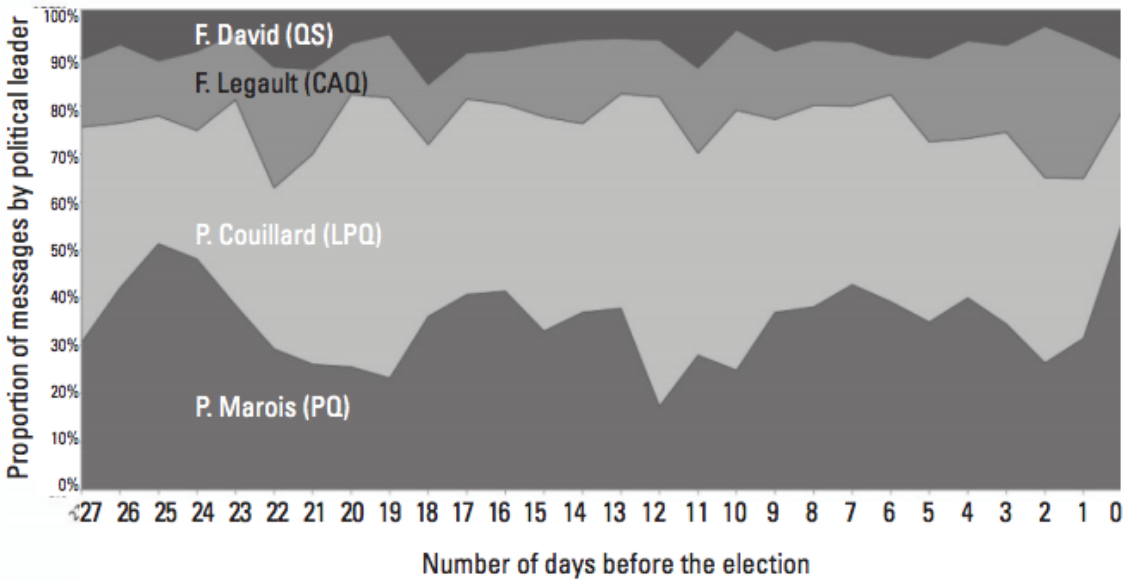

Figure 4.3 Proportion of messages by political leader, 2014

monitoring and feedback tool for political purposes

4.5 Literature review

The research question inscribes this article in a branch of the literature interested in how voters respond to information [Kendall *et al.*, 2015]. This literature includes several relevant empirical contributions, among these are [Ansolabehere *et al.*, 1994, Green et Gerber, 2015], and [Nickerson, 2008].

Launched in 2006, Twitter has since been the subject of prolific academic research. Scholars in political science have benefited from a constant new stream of information from citizens, political parties, and government. Since Twitter data is openly accessible through its API (Application Programming Interface), electoral campaigns can be understood in a different point of view compared to traditional polls. For a systematic literature review on a fragmented field, see [Jungherr, 2016].

The first election of President Obama in 2008 was coined as the first social media election due to the extensive usage of Facebook during the campaign. The United States acted as a "networked nation" [Cogburn et Espinoza-Vasquez, 2011], since millions of users were connected and expressed their opinions regarding policy issues. Such perspective resonates even at the local level, for example the 2010 municipal election in Calgary, Canada, from which [Dumitrica, 2014] has characterized social media as "the new tools and spaces of an improved

communicative relation between politicians and citizens" (p. 65). [Raynauld et Greenberg, 2014] found that Twitter contributes to "permanent campaigning" (p. 413) strategies for political candidates. As mentioned by [Enli et Naper, 2016], the incumbent advantage on social media played in favour of Barack Obama's second election in 2012 against Mitt Romney, since the president leveraged his larger audience of followers to mobilize grassroots actors. In Québec, the National Assembly has adopted a strategy to use internet and social media to promote "an effective communication between the people and their government" (p. 31) [Grétas *et al.*, 2014] since 2009. In fact, online petitions that were introduced since 2009 have collected more signatures from citizens than their paper counterparts ; a Facebook page and Twitter accounts were opened in 2012, whereas a YouTube channel was put in place in 2013.

4.5.1 Extracting information from tweets

Scholars have rapidly tested whether Twitter could be used as a reliable source to predict elections by computing the mean absolute error (MAE) of election forecasts. Results within the margins of error of traditional polls have been obtained (see [Bermingham et Smeaton, 2011, Lui *et al.*, 2011, Tumasjan *et al.*, 2010], with MAEs of 1.65 percent, 1.1 percent, and 5.85 percent respectively).

A tweet, although composed of 140 characters, contains more than 40 elements in its metadata : the name of the user that sent the message, its geolocation (if activated), the time the message was sent, the content of the message, and how many times the message has been liked (previously favourited), to name a few. In this regard, the number of followers a user has could be assessed to measure its reputation or how much attention s/he can generate online. Moreover, metrics such as the sentiment associated with a message or how many times it has been retweeted provide additional information. The later metric could help visualize the network of users concerning a certain topic.

[Bruns et Burgess, 2011] found that during the 2011 Australian federal election, 35 percent of the 415,000 messages in their #ausvotes dataset were retweets (a message sent by a user displayed on another user's feed), whereas 20 percent were messages directly addressed to other users (with the mention @). In South Korea, [Song *et al.*, 2014] associated presidential candidates with specific topics by revealing their most frequently associated terms on Twitter.

Besides a textual analysis of messages on Twitter, several network analyses have been performed during elections [Burgess et Bruns, 2012]. With this scope, [Larsson et Moe, 2013] identified the most prolific Twitter users during the 2011 Danish election using the hashtag #fv11. They provided an answer to who communicated the most in the public sphere (citi-

zens, experts, media, or politicians). The authors offered some interesting insights on citizens' involvement in the public debate. They used network analysis software (Gephi) to present how users are mentioning or retweeting themselves. Such an approach was used in financial conversations to differentiate accounts considered as influential, talkative, or followed regarding the S&P500 stocks [de Marcellis-Warin *et al.*, 2015]. During the 2010 U.S. midterm election, [Conover *et al.*, 2011] analyzed 250,000 tweets to study the polarization of users on the platform. They differentiated two kinds of networks : 1) the retweet network and 2) the mention network. While the first one was highly polarized, with users retweeting messages with whom they agree politically, the second one displayed links between different groups of political affiliation. During the 2011 Canadian election [Gruzd et Roy, 2014] analyzed 5,918 messages sent by 1,492 users on Twitter to understand the communication patterns between politically identified users. In fact, the research suggests the presence of "pockets of political polarization," (p. 28) as observed by the network analysis of the mentions within the dataset. However, cross-ideological connections are also noticed. More precisely, supporters of left-leaning parties (the Liberal Party of Canada, the New Democratic Party of Canada, and the Green Party of Canada) tend to communicate openly between each other, whereas communication toward supporters of the Conservative Party of Canada seems to be more sarcastic and confrontational. By sending tweets using the dedicated hashtag of the election (#elxn41), [Gruzd et Roy, 2014] suggests that "supporters of different parties are aware of each other's presence on Twitter, and that the Twitter communication platform is conducive to exposing people with opposing points of views" (p. 39).

Political debates are considered as key moments in an electoral campaign. Assessing the reaction of television viewers is of great value for political parties since they have the ability to react and frame their messages. By its real-time nature, Twitter could be helpful in doing so. In 2008, the debate between Canadian party leaders was analyzed through comments on Twitter [Elmer, 2013]. In Norway, the 2011 election presented two television debates. Scholars found that Twitter discussions reflected topics opposing candidates on television. However, the social media served as a channel for criticizing the debates, but also for supporting candidates [Kalsnes *et al.*, 2014].

4.5.2 The limits of Twitter as a predictive tool for elections

While the literature offers promising results, a lack of reproducible methods has been noticed. In fact, scholars found that Twitter's ability to predict the outcome of an election may be lower than Facebook's [Cameron *et al.*, 2013]. Concerning the 2011 Singaporean election, [Choy *et al.*, 2012] obtained results as high as 6.06 percent in terms of MAE with only

two candidates. In an article published in 2011, [Gayo Avello *et al.*, 2011] analyzed 234,000 messages during the 2010 U.S. Senate election and obtained an MAE of 17.1 percent using only the number of messages concerning candidates. Their MAE decreased to 7.6 percent when considering the sentiment associated to each message.

[Gayo-Avello, 2012a] resumes concerns regarding the use of Twitter as a predictive tool for elections in "I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper" : A Balanced Survey on Election Prediction Using Twitter Data. More precisely, the author expresses concerns regarding the lack of a balanced literature in the field since most studies present positive correlations between Twitter predictions and electoral outcomes. In addition to this, rumors and propaganda are ignored while the demography on Twitter do not replicate the electoral demography.

In this regard, [Small *et al.*, 2014] studied the propensity of the Canadian population to participate in political debates on the internet. Even though most of the members of parliament are using Twitter (80%), only a fraction of the population is active on social media (3.9% of the overall population of the 2014 Canadian Online Citizenship Survey is following a political actor on Twitter ; 3.1% of the overall population has written a political tweet. Such metrics suggest that Canadian online activity level is lower than the U.S.'s. In this regard, should this low participation number harm the potential of Twitter as a viable source of information regarding politics ? Three elements suggest otherwise. First, the study from [Small *et al.*, 2014] was made outside of a political campaign cycle, when the interest of the population regarding politics is lower than during an electoral campaign. Indeed, during the electoral campaign, news channels and newspapers are offering content dedicated to the campaign, and campaign signs are spread across streets, which would drive discussions about politics. Second, Twitter, as well as other social media, is characterized by the fact that a small proportion of users are driving the discussions by being more vocal than other ones. On the other hand, most of the people are either not participating, or participating a little. This is replicating what would happen in public life with interactions between individuals, where politically engaged people are more vocal than other ones [Barberá et Rivero, 2015, Vaccari *et al.*, 2013]. Finally, even though people are not participating actively by publishing a political tweet, they can be exposed to political content through the accounts they follow. [Vaccari *et al.*, 2013] describe this as the two levels of influence on Twitter during electoral campaigns. Indeed, the authors found that online and offline activities are not distinguished by Twitter users, and that discussions between people offline can be driven by online (Twitter) events.

4.5.3 Québec elections on Twitter

In Québec, a few studies have been published linking Twitter and politics. A comparative study of France and Québec was assessed in 2013 [Eyries et Poirier, 2013] and revealed a later adoption of the social media by Québec's political parties in 2012. In a book published in 2013, [Giasson *et al.*, 2013] describe in detail the 2012 Québec election and in particular how political parties and citizens used Twitter. In his book chapter, [Beauchesne, 2013] sheds light on how that the platform offered a space for political communication unfiltered by traditional media or organizations. By performing a content analysis of almost 1.5 million tweets using a Latent Dirichlet Allocation (LDA) method, political topics emerged from discussions on Twitter. They also found that 54.5 percent of all messages where in fact retweets. In a subsequent chapter, [Giasson *et al.*, 2013] analyzed how each of the six political parties used Twitter in their communication strategies. Two of the smaller ones (Québec Solidaire and Option Nationale) used social media in order to foster interactions with community members, whereas the more established parties (the Québec Liberal Party, the Parti Québécois, and Coalition Avenir Québec) used Twitter in order to broadcast information.

Finally, for a detailed analysis of how Twitter was used during the 2014 Québec election by politicians, see "La cyberdémocratie québécoise : Twitter bashing, #VoteCampus et selfies" [Sullivan et Bélanger, 2016], which studies 13,000 messages sent by 26 "super-users" (candidates from the main parties particularly active on Twitter) and shares the same results as [Beauchesne, 2013] : members of traditional political parties use Twitter as a marketing tool, whereas Québec Solidaire's members promote the use of the social platform to engage with users.

4.6 Methodology

The research question is whether the conversations that took place on Twitter can help us understand the unexpected reversal of the 2014 electoral campaign in Québec. On the one hand, the incumbent party has a tactical advantage by setting up the agenda of the electoral campaign. However, such advantage did not play out well for the Parti Québécois. On the other hand, there is a need to understand how political messages are perceived by the population through social media. At the end of the introduction, five sub-questions were laid out to focus the research question. Let us rephrase them into two working hypotheses to be tested :

— **H1** On Twitter in Québec, certain topics increase the polarization of the public discussion, and their effects last for a longer period of time in comparison to other subjects.

— **H2** On Twitter in Québec, the incumbent advantage plays atthe beginning of the campaign, but can also become a curse toward the end.

4.6.1 Data

This article focuses on the resonance of election topics.In Québec, provincial elections were held from March to April 2014. Tweets regarding each party leader and the electoral campaign were collected. More than 670,000 messages were analyzed in order to create the sample. Messages on Twitter were collected that presented at least one of the three used hashtags during the election period, namely #assnat (National Assembly), #polqc (Québec politics), and #qc2014 (Québec 2014). In total, this dataset was composed of 672,497 messages from the REST API of Twitter using [Barbera, 2018] streamR package on R.

Each day, different indicators of importance on Twitter were computed. First, the volume of tweets corresponding to each party leader was calculated (Pauline Marois for the Parti Québécois, Philippe Couillard for the Quebec Liberal Party, François Legault for the Coalition Avenir Québec, and Françoise David for Québec Solidaire). Concerning Québec Solidaire, only Françoise David was considered since she was the most known political figure compared to the second spokesperson of the party, Andrés Fontecilla. Second, the presence of each party was assessed on a daily basis through their respective hashtags (#PQ, #PLQ, #CAQ, and #QS). Finally, the goal was to capture how the campaign topics would evolve and what would be the response from the population on Twitter.

To explore this, the four electoral programs as put forward by the four parties were consulted. 31 ideas or concepts that were going to be promoted during the campaign by the four leaders were isolated. Out of the 670,000 messages, a dataset of 157,916 tweets mentioning one or more of the 31 concepts was assembled. In total, this dataset is made of 868 observations per variable of interest (31 concepts x 28 days). Those 31 concepts were organized in four general categories regarding the independence of Québec, ethics, economics, and society. The volume of messages corresponding to each main category was obtained by assessing how many messages were written about their corresponding keywords. For the independence of Québec, messages containing words related to the Québec Charter of Values, national identity, secularism, referendum, and sovereignty were considered. Words related to the Charbonneau Commission, collusion, corruption, ethics, and integrity were searched for ethical behaviours. Finally, unemployment, debt, economy, employment, taxation, federal and provincial taxes, taxes, infrastructures, resources, and investments were used for assessing economic topics ; education, students, environment, family, day care, youth, doctors, retirement, health, and university were considered social topics. Those keywords appeared in the electoral programs of

the political parties and were also reflected in discussions in newspaper articles and television debates. In order to evaluate how political parties or political leaders were related to these four main categories, the number of messages associated to each keyword was computed (see Figure **??**).

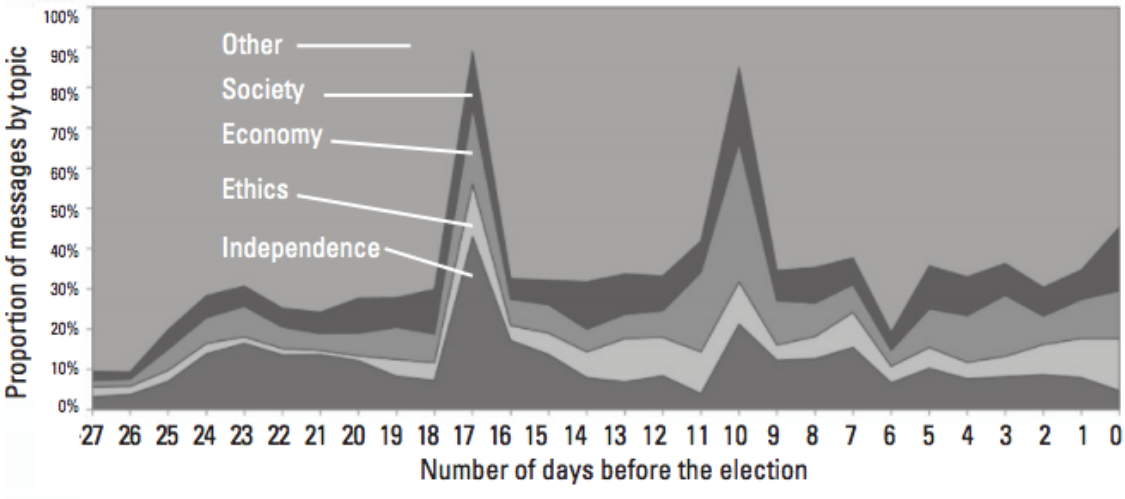

Figure 4.4 Proportion of messages by topic, 2014

Finally, the electoral campaign was divided in three separate time periods. These three periods are delimited by the two televised debates in order to analyze how perceptions have changed during the electoral campaign. The first period finished the day before the first debate (March 17, 2014), the second period starts on the day of the first debate until the day before the second debate (March 18–26, 2014), and the third period goes from the second debate until the end of the electoral campaign (April 7, 2014) (see Figure **??**).

For the topic of Independence, the Québec Charter of Values was the keyword that generated the highest number of tweets. Discussions on ethical behaviours were dominated by messages about integrity. About the economy, both employment and the economy in general emerged as prominent subjects. Finally, health was the main concern of the messages written about society. Figure 4.6 presents how each of the four categories have generated discussions on Twitter.

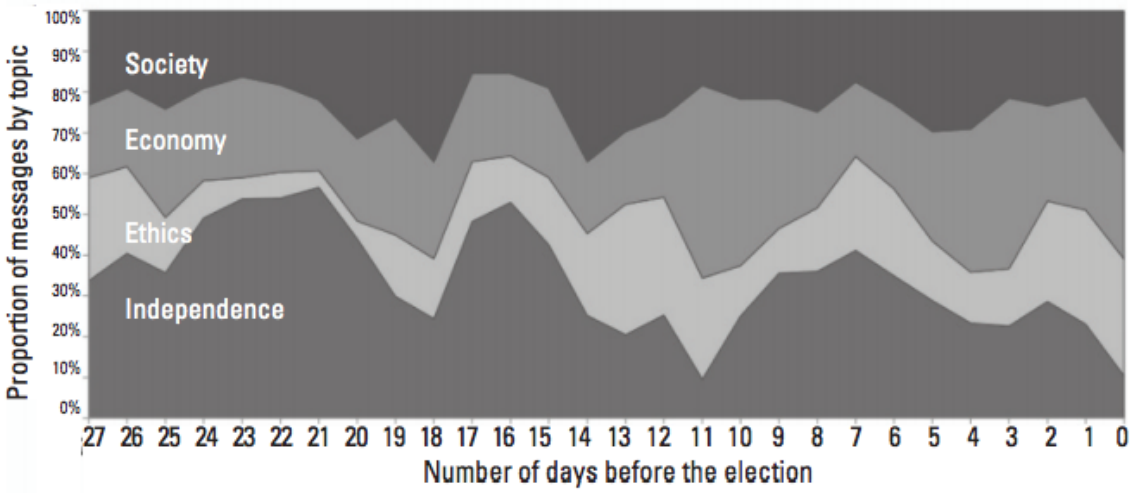

Figure 4.5 Proportion of messages by topic, 2014

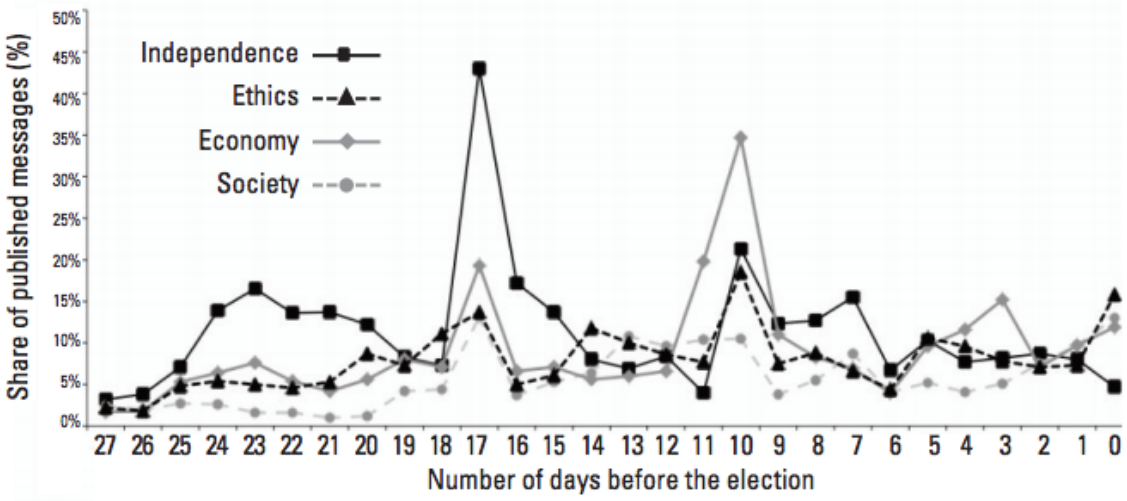

Figure 4.6 Share of published messages in percentage

Table 4.1 Descriptive statistics about the dataset for each category studied

	Min.	Max.	Average	Median	Std. Dev.
Topic 1: Independence	491	7174	2576.11	2047	1465.99
Independence	76	507	206.36	181.5	107.46
National Identity	0	273	64.61	29	74.84
Québec Charter of Values	126	3573	1226.18	892	862.86
Referendum	78	2659	616.79	467.5	539.41
Secularism	16	315	121.39	94	82.19
Sovereignty	109	912	340.79	312.5	176.43
Topic 2: Ethics	147	2012	875.61	727.5	560.01
Charbonneau Commision	1	779	102.18	52.5	154.92
Collusion	0	61	25.68	24.5	16.71
Corruption	77	778	289.93	383	180.29
Ethics	8	558	117.46	77	119.36
Integrity	29	885	340.36	274.5	254.48

4.6.2 Model

The 157,916 tweets were put in a cross-section time-series format. The sections were structured around the 31 keywords and through the 28 days during which the tweets were collected.

Two models are used to measure 1) the importance of each topic during the campaign and 2) how likely a party was to be associated with such topics throughout the campaign.

For the first model, the dependent variable is a categorical variable capturing whether a message belongs to the first category (Independence, 1), the second category (Ethics, 2), the third category (Economy, 3), or the fourth category (Society, 4). The presence of the political parties in the discussion about the keywords constituting each category is used as independent variables. The controlled variables are the three time periods identified. Finally, an ordered logistic model is used as follows :

$$pr(y_t) = x_1\alpha_t + x_2\beta_t + x_3\gamma_t + x_4\delta_4 + x_5\tau_5 + x_6\tau_6 + c \tag{4.1}$$

Where $\{\alpha_t; \beta_t; \gamma_t; \delta_t\}$ equals the number of tweets per day referring to the (Parti Québécois ; Quebec Liberal Party ; Coalition Avenir Québec ; Québec Solidaire) and τ_i the period of the campaign studied, with $i = \{2; 3\}$ for the second and third period (with the first period as reference).

For the second model, the party ahead in terms of mentions for each keyword was computed on a daily basis. If two parties were equally mentioned for a given day, the observation was duplicated in order to account for each party. In this model, a binary dependent variable was used concerning each topic, such as : $y_{t,\lambda} = \{1; 0\}$, with $y_{t,\lambda} = 1$ when observing for a topic $\lambda = \{1; 2; 3; 4\}$ and 0 otherwise.

As independent variables, two categorical variables were considered (the party leading in terms of mentions for a given day and the period ofthe campaign). A logistic model was considered as follows :

$$pr(y_{t,\lambda}) = x_7\pi_t + x_8\tau_8 + c \tag{4.2}$$

With $\pi_t = \{1; 2; 3; 4\}$ for the leading party in terms of mentions for a given day, referring to the (Parti Québécois ; Liberal Party of Québec ; Coalition Avenir Québec ; Québec Solidaire) and $\tau_t = \{1; 2; 3\}$ concerning each period of the campaign.

4.7 Results

4.7.1 Most important topic of the campaign

Throughout the whole period, the probability that the tweets mentioning the four parties are about "Independence," "Ethics," "Economy," or "Society" was 17.8 percent, 20.1 percent, 35.4 percent, and 26.6 percent respectively.

Therefore, the most prevalent topic during the whole electoral campaign in the dataset, when mentioning the different political parties, was the "Economy" and then "Society." It is interesting to note that the Parti Québécois thought that questions of identity and independence were going to be important during this election. Apparently, they were not as important as assumed.

4.7.2 Party association during the campaign

To go further in this analysis, a breakdown of estimations by party and time periods is needed. Tables 4.4 to 4.7 present the results of the predicted probability concerning each topic using the second model.

Two parties are more associated with the topic of "Independence" (a statistically significant relation) during the campaign. The PQ is the more prevalent one, since 28.8 percent of its campaign is associated with this topic. On the other hand, 2.5 percent of the PLQ campaign is associated with this topic when the party is leading in terms of mentions. There is a decline in the association between "Independence" and the PQ, especially at the end of the campaign (from 28.8% to 26.7%). This is particularly interesting when one considers that the PQ — the incumbent government — favoured this category when deciding to launch a new election.

The QS was never a prominent figure regarding "Ethics" compared to the other parties during the campaign. This time, the CAQ was associated to this second topic, but the relation is still not statistically significant. The PLQ, when leading in terms of mentions, was on average 34.5 percent of the time associated to the topic of "Ethics."

Concerning the "Economy," all parties have been leading the conversation during the campaign, but to a different extent. More precisely, most of the CAQ's and QS's campaigns were associated with this topic (more than 53% on average). It is interesting to note that the economic topics were the ones put forward by the CAQ. Although they did not resonate at the beginning of the campaign as much as the category, "Independence," when the economy became important for the users on Twitter, then the CAQ made a comeback (see Figure 4.3).

Finally, when leading in terms of mentions on Twitter concerning the "Society" topic, the

order of the most associated parties goes as follow : CAQ, QS, PQ, and PLQ. In conclusion, CAQ and QS did a better job than the other parties to be more associated with this category and also to help change the initial agenda. The results are summarized in Figure 4.7.

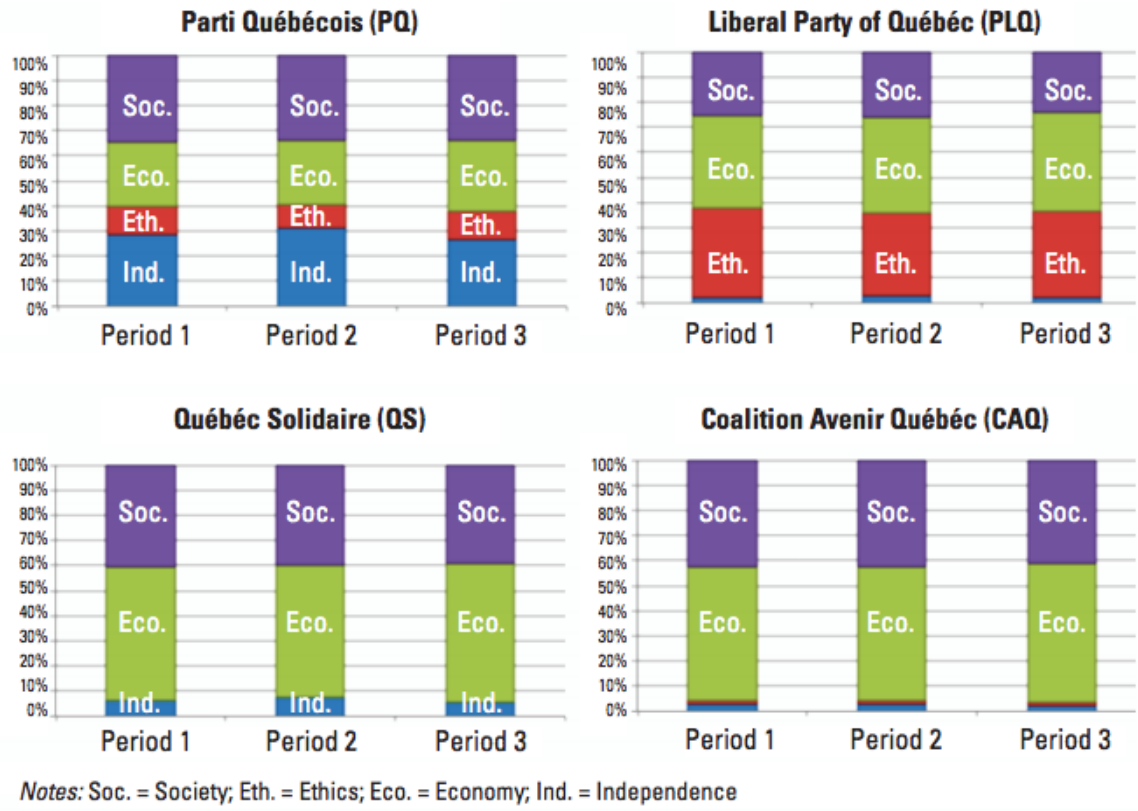

Figure 4.7 Election results summarized

4.8 Conclusion

During the 2014 Québec Election, the gamble made by the PQ did not pay off well. While being in power before calling the election, the PQ lost to the PLQ, which won a majority government a month later. In this article, discussions held on Twitter regarding three hashtags used during the campaign (#assnat, #polqc, and #qc2014) were collected and analyzed. After scrapping 670,000 messages, the goal of the article was to understand how political parties were perceived during the electoral campaign by Twitter users.

While the incumbent government pushed forward the topic of "Independence" at the start of the campaign, it did not last too long and was surpassed by other events during the campaign, especially discussions about "Economy" and "Society," which were put forward by other parties. This evolution in association between four main topics (Independence, Economy,

Society, Ethics) and the four main parties (PQ, PLQ, QS, CAQ) is the main contribution of this article. When leading in terms of mentions, each party was quantitatively associated with a different topic, and this association evolved during a campaign that has been paced by two televised debates.

A potential lesson is that the incumbent advantage plays at the beginning of the campaign, but can also become a curse toward the end. Parties need to anticipate what will be the main topic right before election day, and not so much at the beginning of the campaign. In terms of political strategy, opposition parties to the incumbent government should not fall into the trap of spending time on the incumbent agenda.

This article helped characterize which topic each party was associated with during the last month of the election. The results could be used to implement a framework to assess the perception of the political parties on social media. In the Québec context, this article provides information on a still-limited research area (social media in Québec politics) within a growing research field (politics and social media). Concerning political parties, such methodology may be useful to gauge their political propositions and how Twitter users are responding to it. For polling agencies, such a framework would provide additional information to traditional polling methods.

As mentioned in the literature review, limitations do exist such as the pitfall of using direct social media information as a substitute for voting intentions (hence having sample biases). Further works should take into account the dynamics of Twitter interactions (i.e., who are the most influential individuals concerning political discussions ?) or topic differences between types of individuals (do age or gender influence the type of discussions or associations ?). Finally, a growing literature on political bots (automated accounts) is emerging, and should be kept in mind for the next studies on elections, such as the one in 2018 in Québec.

Table 4.2 Descriptive statistics about the dataset for each category studied (continued)

	Min.	Max.	Average	Median	Std. Dev.
Topic 3: Economy	265	5307	1388.5	1084	1015.29
Debt	6	491	153.86	117	120.57
Economy	108	973	377.46	367.5	178.48
Employment	48	1159	333.64	247.5	301.12
Federal and Provincial Taxes	8	781	143.54	86.5	170.8
Fiscal Policy	2	1674	153.36	31	360.6
Infrastructures	0	111	18.11	10.5	24.5
Investments	18	368	75	55.5	67.93
Resources	0	34	10.46	9	8.11
Taxes	9	461	96.29	76	86.04
Unemployment	2	217	26.79	15.5	42.8
Topic 4: Society	279	2833	1219.89	1136	581.1
Day Care	21	301	81.57	60	68.53
Doctors	5	363	81.04	51.5	78.15
Education	7	323	101.54	88.5	70.38
Environment	24	784	139.82	101.5	145.28
Family	13	457	131.04	110.5	99.27
Health	77	850	285.07	220.5	176.33
Retirement	1	122	25.39	16	31.41
Students	6	890	200	163.5	176.69
University	4	166	40.39	28.5	36.44
Youth	15	298	134.04	119.5	68.9

Table 4.3 Predicted probabilities for each category based on an ordered logit estimation

Pr(Independence)					
	Margin	Std. Err.	*P*-value	[95% Conf. Interval]	
constant	.1784479	.0142735	***	.1504724	.2064235
N = 868					
P-value: *** < .01; ** < .05; * < .1					
Pr(Ethics)					
	Margin	Std. Err.	P-value	[95% Conf. Interval]	
contant	.2008826	.0156595	***	.1701905	.2315748
N = 868					
P-value: *** < .01; ** < .05; * < .1					
Pr(Economy)					
	Margin	Std. Err.	*P*-value	[95% Conf. Interval]	
constant	.3544883	.017434	***	.3203183	.3886582
N = 868					
P-value: *** < .01; ** < .05; * < .1					
Pr(Society)					
	Margin	Std. Err.	*P*-value	[95% Conf. Interval]	
constant	.2661812	.158944	***	.2350286	.2973337
N = 868					
P-value: *** < .01; ** < .05; * < .1					

Note: Predicted probabilities for each category based on an ordered logit estimation.

Table 4.4 Predicted probabilities for the category "Independence"

Pr(Independence)					
Period – Pol. Party	Margin	Std. Err.	*P-value*	[95% Conf. Interval]	
1 – PQ	.2883107	.0356576	***	.2184232	.3581983
1 – PLQ	.0246657	.0106295	**	.0038322	.0454991
1 – QS	.0652011	.0453452		-.0236738	.154076
1 – CAQ	.0285819	.0201166		-.0108459	.0680098
2 – PQ	.3162395	.0358472	***	.2459803	.3864988
2 – PLQ	.0280621	.0116719	**	.0051855	.0509387
2 – QS	.073757	.0506823		-.0255785	.1730925
2 – CAQ	.0324996	.0230377		-.0126535	.0776528
3 – PQ	.2667128	.0281657	***	.211509	.3219165
3 – PLQ	.0222018	.0093918	**	.0037943	.0406093
3 – QS	.0589327	.0413546		-.0221207	.1399861
3 – CAQ	.0257372	.0183988		-.0103239	.0617982
N = 885, *P*-VALUE: *** <.01; ** <.05; * <.1					

Table 4.5 Predicted probabilities for the category "Ethics"

Pr(Ethics)					
Period – Pol. Party	Margin	Std. Err.	*P-value*	[95% Conf.	Interval]
1 – PQ	.1094821	.0204327	***	.0694348	.1495294
1 – PLQ	.3588622	.0492978	***	.2622403	.4554842
1 – CAQ	.0151309	.0150855		-.0144363	.044698
2 – PQ	.0959016	.0179669	***	.0606871	.1311162
2 – PLQ	.3256603	.0411761	***	.2449567	.4063639
2 – CAQ	.0130821	.0131888		-.0127676	.0389317
3 – PQ	.1063555	.0170065	***	.0730235	.1396876
3 – PLQ	.3514244	.0423167	***	.2684851	.4343636
3 – CAQ	.0146544	.0147334	***	-.0142225	.0435313
N = 885, *P*-VALUE: *** <.01; ** <.05; * <.1					

Table 4.6 Predicted probabilities for the category "Economy"

Pr(Economy)					
Period – Pol. Party	Margin	Std. Err.	*P-value*	[95% Conf.	Interval]
1 – PQ	.255428	.0292545	***	.1980902	.3127658
1 – PLQ	.3686906	.0424727	***	.2854456	.4519355
1 – QS	.5238003	.0949607	***	.3376808	.7099199
1 – CAQ	.5285858	.0632392	***	.4045805	.6525911
2 – PQ	.2577098	.0285069	***	.2018372	.3135823
2 – PLQ	.3714793	.0383998	***	.2962171	.4467416
2 – QS	.5267833	.0947099	***	.3411554	.7124113
2 – CAQ	.5315657	.067472	***	.399323	.6638083
3 – PQ	.2804084	.0257722	***	.2298958	.3309211
3 – PLQ	.3988148	.039319	***	.321751	.4758786
3 – QS	.5554475	.0944638	***	.3703017	.7405932
3 – CAQ	.560182	.0666578	***	.4295352	.6908288
N = 885, *P*-VALUE: *** <.01; ** <.05; * <.1					

Table 4.7 Predicted probabilities for the category "Society"

Pr(Society)					
Period – Pol. Party	Margin	Std. Err.	*P-value*	[95% Conf. Interval]	
1 – PQ	.3460254	.0331725	***	.2810085	.4110423
1 – PLQ	.2542577	.0364651	***	.1827874	.325728
1 – QS	.403192	.0929967	***	.2209218	.5854622
1 – CAQ	.4228356	.0624102	***	.300514	.5451573
2 – PQ	.3437164	.0322691	***	.2804701	.4069628
2 – PLQ	.2523248	.0333266	***	.1870058	.3176438
2 – QS	.4007353	.0926896	***	.2190671	.5824035
2 – CAQ	.4203436	.0663559	***	.2902883	.5503988
3 – PQ	.3379618	.0277283	***	.2836153	.3923083
3 – PLQ	.2475232	.0330615	***	.1827239	.3123226
3 – QS	.3946	.0928149	***	.2126863	.5765138
3 – CAQ	.4141155	.0659817	***	.2847939	.5434372

N = 885, *P*-VALUE: *** <.01; ** <.05; * <.1

CHAPITRE 5 ARTICLE 2: THE 2015 CANADIAN ELECTION ON TWITTER: A TIDY ALGORITHMIC ANALYSIS

5.1 Présentation de l'article

Référence. Sanger, W. et Warin, T. 2017. The 2015 Canadian Election on Twitter : A Tidy Algorithmic Analysis. Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence. IEEE Xplore. pp. 910 - 915.

Le second article de la thèse porte sur la campagne électorale canadienne de 2015. Alors troisième candidat selon les intentions de votes au début de la campagne, Justin Trudeau réussit le tour de force de former un gouvernement majoritaire. Stephen Harper, Premier ministre sortant, n'est pas réélu ; Thomas Mulcair, premier dans les sondages au début de l'élection, termine en troisième position. Comment les chef.fe.s de partis politiques sont perçu.e.s durant la campagne électorale sur les médias sociaux ?

Cet article puise dans plus de 3,5 millions de messages publiés sur Twitter pour répondre à cette question de recherche. Deux méthodologies quantitatives d'analyse de texte ont été notamment utilisées. Une analyse de sentiment permet de mettre en évidence la teneur du contenu lexical concernant chaque politicien·ne tandis qu'une analyse par allocation de Dirichlet latente (LDA) regroupe les thématiques associées aux politicien·ne·s.

Cet article est coécrit avec Thierry Warin a été accepté pour publication à l'automne 2018 sur IEEE Explore. Il a été mis sous presse au printemps 2018 et fait suite au 4$^{\text{th}}$ International Conference on Computational Science & Computational Intelligence (CSCI) de Las Vegas aux États-Unis.

Des versions préliminaires ont fait l'objet de plusieurs présentations orales, dont :

— Sanger, W. et Warin, T. (2017). The 2015 Canadian Election on Twitter : A Tidy Algorithmic Analysis. 4$^{\text{th}}$ International Conference on Computational Science & Computational Intelligence (CSCI). Décembre 2017, Las Vegas, États-Unis.
— Sanger, W. et Warin, T. (2016). Résonance de propositions politiques sur Twitter : le cas de la campagne électorale canadienne de 2015. ACFAS, 9-13 mai 2016, Montréal, Canada.
— Sanger, W. et Warin. T. (2016). The 2015 Canadian Election : Twitter to Measure the Public's Reaction during Televised Debates. Intelligences numériques. 4-6 mars 2016, Québec, Canada.
— Sanger, W. et Warin, T. (2015). Analyser des données massives sur les réseaux so-

ciaux, le cas de la campagne électorale canadienne sur Twitter. LabCMO (UQAM), 6 novembre 2015, Montréal, Canada.

— Sanger, W. et Warin, T. (2015). Démocratie au XXI^{e} siècle : les technologies au service des citoyens. CIRANO, 27 octobre 2015, Montréal, Canada.

5.2 Abstract

During the 2015 General Election in Canada, the Liberal Party of Canada was elected to form a majority government, despite being third at the start of the electoral campaign. How was perceived the incumbent party and its rivals during the election ? By using a tidy approach of a massive dataset collected on Twitter (3.5 millions tweets), we developed two methodologies to characterize how politicians were perceived on social media during the election. First, a sentiment analysis was performed regarding each political leader, then by using the whole dataset, different topics of the election were associated to each leader through a Latent Dirichlet Allocation (LDA).

Keywords. Social Media Analysis, Elections, LDA, Social Data Science, Canada

5.3 Introduction

Recent elections (United States, 2016 ; Brexit referendum, 2016) have illustrated the dynamics of information generation on social media. Overwhelmed with "algorithmic propaganda", voters that are exposed to fake news may go to the polling stations with false information in mind, or perceive their representative through distorted lenses. Incumbent candidates might have to react to disinformation in addition to pursue their usual electoral campaign.

In 2015, a General Election occurred in Canada. In this case, the incumbent party was not reelected (the Conservative Party of Canada). From being third in the polls at the start of the electoral campaign, the Liberal Party of Canada and its leader, Justin Trudeau, obtained a majority victory. The Canadian political system is a representative democracy in which citizens elect their representatives in the Canadian Parliament. The election cycles are about 4-year long, but could be shortened by the governing party or by a coalition of opposing parties, leading to the well-known hypothesis in Political Science about the incumbent advantage. During each election, candidates from political parties are asking voters from electoral districts to vote for them. As part of a first-past-the-post voting system, a political candidate will become a member of the Canadian Parliament if s.he succeeds in securing one more vote than the second best candidate.

Traditionally, few main parties are in competition during an election cycle. As of the beginning of 2017, 18 official parties are registered at the federal level. However, few of them are effectively represented in the Parliament and in the media. At the federal level, five main parties were in competition in 2015, namely the Liberal Party of Canada (lead by Justin Trudeau), the Bloc Quebecois (lead by Gilles Duceppe), the Green Party (lead by Elizabeth May), the Conservative Party of Canada (lead by former Prime Minister Stephen Harper) and New Democratic Party of Canada (lead by Thomas Mulcair).

In this paper, we use a unique dataset gathered during the electoral campaign in 2015 from which we extract structural differences regarding each candidate. How was perceived each candidate during the electoral campaign ? In Section 2, we present a comprehensive review of the literature on social media and politics. Section 3 will focus on the different methodologies employed, from the Data Science platform assembled to the structure of the dataset and both linguistic methods. Results are presented in Section 4 while we will bring some perspectives and highlight the limitations of our paper in the last section.

To the best of our knowledge, this is one of the first contributions bridging Computational Science, Political Science, while being applied to Canadian elections.

5.4 Literature Review

5.4.1 Systematic Literature Review

Using Twitter to understand electoral dynamics has fostered a prolific stream of research. In fact, as available on Web of Science, 7,607 documents (academic papers, books, conference proceedings,...) refer to the topic of social media and politics. More precisely, on September 29 th 2017, we have selected the terms "social media" or "Twitter" in the topic search box of the academic aggregator, as well as the regular expressions referring to "election", "electoral" or "politic" as a second topic of research. The years considered for this research are from 2005 until 2017, hence excluding results prior Twitter's creation.

By using VOSviewer, a software developped by van Eck and Waltman to analyze bibliometric data, it is possible to produce a comprehensive systematic review of the field of social media (and Twitter) applied to political science, something a human being could not do without computational power. From our dataset of 7,607 documents, we selected the keywords that are shared by at least 7 documents. We then omitted keywords that were capturing the overall topics of research (i.e. : social media, media, politics, Twitter, Facebook, social networks). We replicated the methodology used by [van Eck *et al.*, 2006] in order to produce Fig. **??**.

5.4.2 Using Algorithmic Techniques with Twitter Data

The study of Twitter has presented promising results regarding electoral predictions. However, methodological concerns have also been raised by scholars. The lack of reproducibility [Chung et Mustafaraj, 2011], low predictive power compared to other social network sites such as Facebook [Cameron *et al.*, 2013] and inaccuracy in results [Gayo-Avello, 2012a] are amongst the potential pitfalls [Gayo Avello *et al.*, 2011] of using Twitter as a predictive tool.

This is precisely the reason why we do not want to explain the results of the 2015 Canadian election directly through Twitter. Instead, we want to understand how a population (not a sample), the users of Twitter, reacted concerning their political figures during a period of extreme political exposure for the population of Canada. To do so, we use a set of algorithmic methodologies to insure the replicability of our results, which is traditionnaly a challenge in social sciences despite the multiplicity and availability of data [King, 2011].

It is not the first time such techniques are used on studies bridging politics and Twitter, especially in Canada. In fact, [Beauchesne, 2013] perform a Latent Dirichlet Allocation method (LDA) to reveal political topics during the 2012 election in Quebec. [Song *et al.*, 2014] revealed the most associated terms regarding presidential candidates in South Korea, while

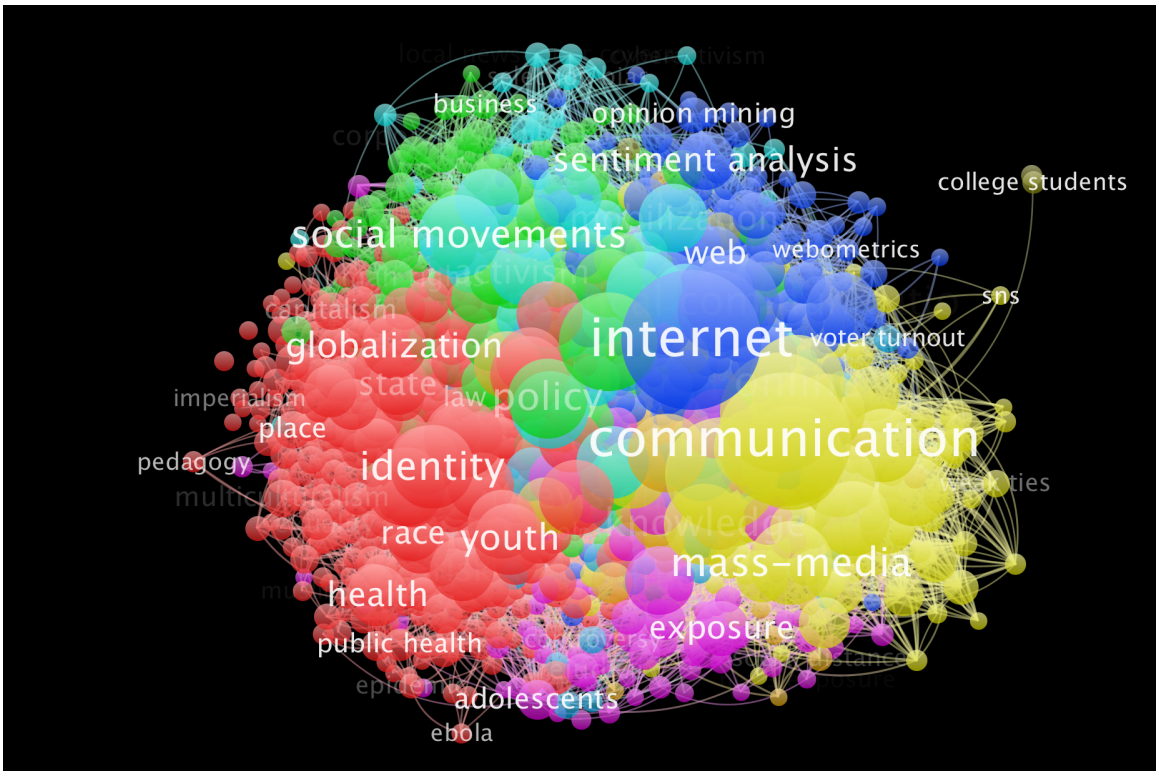

Figure 5.1 Cartography of the keywords extracted from the systematic literature review

in Norway, Twitter discussions reflected the topics opposing candidates during televised debates [Kalsnes *et al.*, 2014]. [Jungherr, 2016] presents a comprehensive review of the literature regarding the use of Twitter in electoral context.

5.4.3 Research Question

How each political leaders was perceived during the 2015 electoral campaign in Canada ? To answer such a question, one popular approach in Social Sciences is the analysis of the content of Twitter messages. However, selecting specific keywords might induce biases in the results. In our research, we use two complementary linguistic computational methods : first a supervised method by associating a sentiment score to each candidate during the electoral campaign, then an unsupervised (semi-supervised) method to highlight which topics are preferentially associated to each candidate.

5.5 Methodology

We assembled a computing platform dedicated to Social Data Science research in order to tackle our research question. This platform (called Nüance-R) consists in servers loaded with an R compiler, MongoDB, Ubuntu 16.04 LTS and several other pieces of software and packages. It is able to collect data from APIs, create visualizations and perform econometrics and data analytics through RStudio's IDE.

5.5.1 Data

In this study, we have extracted messages related to the Canadian election based on a framing strategy [Sanger et Warin, 2018c]. We access the Twitter REST API with the streameR R package [Barbera, 2018] and selected three official hashtags of the election as filters (« #exln42 », « #cdnpoli », « #polcan »). These hashtags were used by political parties, political figures, journalists, television networks, individuals, and were similar to the latest federal election in Canada (except for the first hashtag that was #elxn41, as for the constitution of the 41$^{\text{th}}$ Parliament of Canada). « #cdnpoli » refers to Canadian Politics on Twitter whereas « #polcan » is the French equivalent, Politique canadienne. In total, we considered 3,498,633 tweets published by 218,255 unique users.

Tweets are collected from the first day of the electoral campaign (August, 2$^{\text{nd}}$ of 2015) until one before the end of the electoral campaign (October, 18$^{\text{th}}$ of 2015) for two reasons : (1) first to avoid capturing messages announcing the results of the election and (2) political parties campaigning during the Election day is not legal in Canada. In Fig. 5.2, we describe the total amount of messages collected. From August until the first week of September, the average number of messages collected using one of the three official hashtags was under 38,000 daily tweets (except for August 7$^{\text{th}}$ of 2015 during the first television debate). This average number of collected messages doubled throughout the campaign.

From this initial database of tweets mentioning one of the three official hashtags of the election, we dedicated our analysis on how political leaders were perceived. Hence, in order to select a message concerning each political figure, we searched for regular expressions concerning themselves inside the message portion of a tweet. For four out of the five political leaders, the regular expression referred to their surname (which also was used in their Twitter handle). However for Elizabeth May (Green Party of Canada), in order to not confuse with any word combination containing "may", we used a more complex regular expression referring also to her first name as well as declinations of her first name. The complete list of regular expressions and descriptive statistics is provided in table 5.1. The total number of tweets

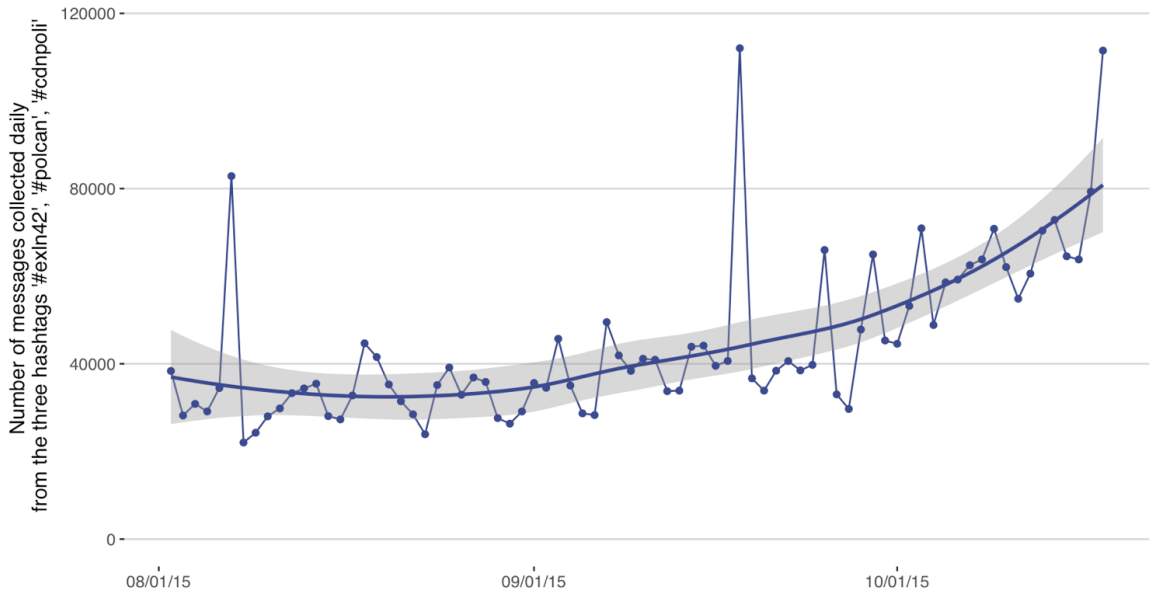

Figure 5.2 Evolution of the number of messages during the campaign

used for this paper is about 1.5 million messages

Table 5.1 Number of topics mentioning by term

Political leader	Regular expression	Number of messages	Number of unique users	Average number of messages per user
Gilles Duceppe	"[Dd]uceppe"	21,190	5,482	3.87
Stephen Harper	"[Hh]arper"	878,365	17,933	48.98
Elizabeth May	"[Ee]li[zs]abeth [Mm][Mma][ya]"	71,722	34,526	2.08
Thomas Mulcair	"[Mm]ulcair"	220,492	84,217	2.62
Justin Trudeau	"[Tt]rudeau"	317,102	47,967	6.61

Political leaders were not mentioned as equally throughout the campaign. In fact, the former Prime Minister Stephen Harper (PCC) was the most mentioned with more than double the amount of messages compared to the second most mentioned political leader (i.e. Justin Trudeau) except during the days with televised debates. Thomas Mulcair was the most tweeted candidate by unique users (84,217), which is 75.57% more important than the number of users tweeting about Justin Trudeau and 369.62% more important than for Stephen Harper. See Fig. 5.3 for more details on the evolution of messages related to each political leader.

Such disparity between mentions of political leaders raise questions about the content of

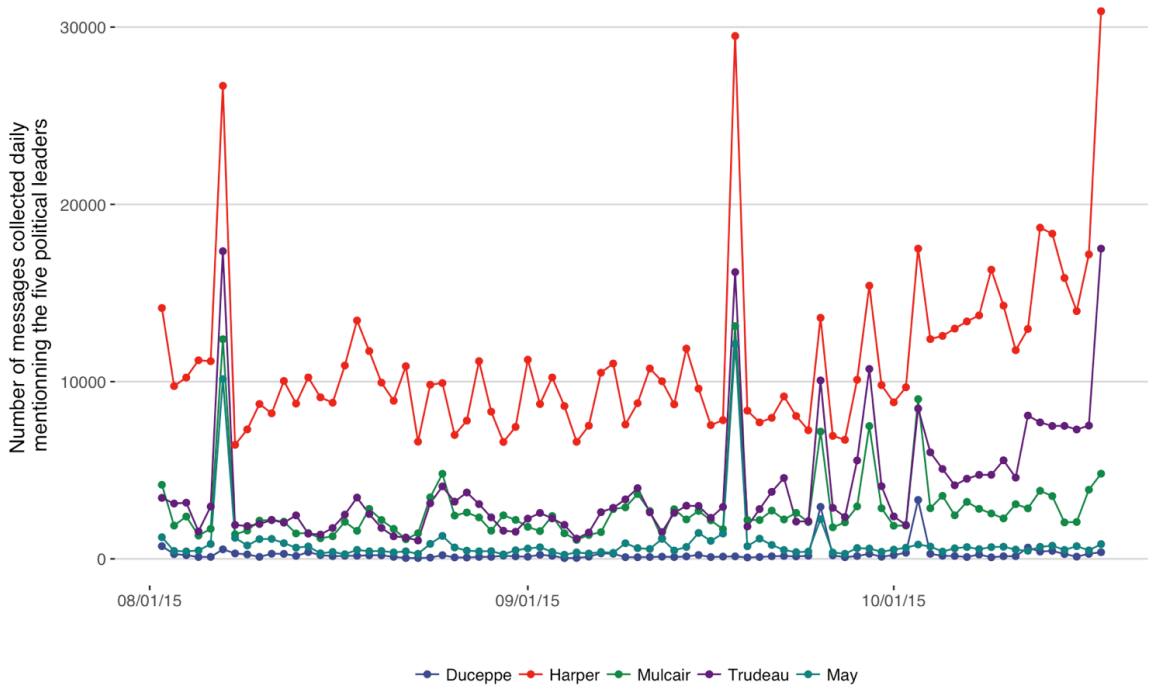

Figure 5.3 Evolution of messages regarding each political leader

these messages. More importantly, why does former Prime Minister Harper generated a ratio of messages per user way more important than his contestors ? What was the content of such messages ?

5.6 Methods

In the following section, we describe two methodologies used to delve into the database of tweets regarding each political leader : (1) a content analysis based on a framing strategy and (2) an unsupervised classification based on a Latent Dirichlet Allocation (LDA) approach.

5.6.1 Content Analysis

The first step of the analytical analysis is to tidy our dataset following Hadley Wickham's description [Wickham, 2014] : "each variable is a column, each observation is a row, each type of observational unit is a table". In order to associate a sentiment score to each tweet we manipulate our dataset in order to remove all links from the messages, then tokenize each message and finally we remove all stopwords following [Silge et Robinson, 2017] approach.

Our framing strategy consists in separating all tweets regarding a political leader into a

dedicated dataset. Then, two lexical databases were used to evaluate the content of each set of tweets. These lexical databases are unigram-based. First, the lexicon of 6,788 words from [Hu et Liu, 2004] assign a binary value of positivity/negativity to each word. A net value of sentiment is assessed by computing the difference between positive and negative words, such as for each day $i = \{1; ...; 78\}$ of the electoral campaign and for each political leader $j = \{Duceppe|Harper|May|Mulcair|Trudeau\}$:

$$sentiment_{Hu.and.Liu;i;j} = positive_{i;j} - negative_{i;j} \tag{5.1}$$

The second lexicon of 2,476 words is from [Nielsen, 2011], which ranks words on a discrete scale from -5 (negative) to +5 (positive). This would provide a finer interpretation of words associated to each political leader since we can deduce how strongly positive or strongly negative words impact each individual by their frequency. Hence, we compute for each political leader the most impactful k words throughout the political campaign such as :

$$sentiment_{Nielsen;j} = SentimentScore_{j;k} * NumberOfOccurence_{j;k} \tag{5.2}$$

5.6.2 Topic Associated to each Political Leader : LDA Analysis

The second step of our methodology is to perform a Latent Dirichlet Allocation on the whole dataset. The goal of this method is to clusterize the dataset in different topics throughout the campaign with an unsupervised learning method. Those topics are then associated to each political leaders to reveal the dominant ideas and conversations.

To do so, the dataset of all tweets is filtered in order to keep words that are found at least more than 50 times. Then, by using the VEM algorithm (variational expected-maximisation, [Blei *et al.*, 2003]) from the topicmodels library in R by [Hornik et Grün, 2011], the dataset is structured into 10 different topics. Finally, we provide the most important words of the categories associated with each political leaders.

5.7 Results

5.7.1 Framed Strategy with Sentiment Analysis

In Fig. 5.4, we compare the results of the sentiment analysis of the messages associated to each political leader throughout the campaign with the lexicon [Hu et Liu, 2004]. Several elements should be noticed. First, while for most candidates the fluctuations in sentiment

have occured during the campaign, the sentiment score remained mostly negative towards the incumbent. Second, not all candidates are perceived the same : Justin Trudeau generated more important fluctuations in the sentiment score than his rivals, with only Thomas Mulcair to be compared to. Third, Gilles Ducepe and Elizabeth May obtained relatively lower absolute values of sentiments, corresponding to a smaller exposure during the campaign. Finally, we can deduce how each candidate were perceived during the tevised debates. For the first debate (August 8th), Trudeau, May, Mulcair and Harper were perceived from the most positive to the least ; for the second debate (September 17th), Mulcair, Trudeau and Harper ; for the third debate (September 24th), May, Duceppe, Mulcair, Trudeau and Harper ; for the second-to-last debate (September 28th), Trudeau, Mulcair and Harper and finally, for the last debate (October 2nd), Duceppe, Trudeau, Mulcair and Harper.

Figure 5.4 Sentiment score concerning each political leader

In figure 5.5, we can highlight from the second lexicon [Nielsen, 2011] the words that contributed the most to each political leader's sentiment score. Here, Stephen Harper is perceived more negatively than his rivals. Justin Trudeau have been more associated to positive concepts than the other political leaders. For example, the 12 words that contributed the most to sentiment scores regarding Justin Trudeau were "support", "win", "amazing", "care", "strong", "promises", "endorses", "won", "wow", "love", "hope" and "fear", all positive words except

the last one. On the contrary, Stephen Harper was associated to only two positive words, “support” and “win”, whereas negative words impacted more his sentiment scores (“scandal”, “fear”, “bad”, “crisis”, “angry”, “lost”, “recession”, “racist”, “war”, “worst”).

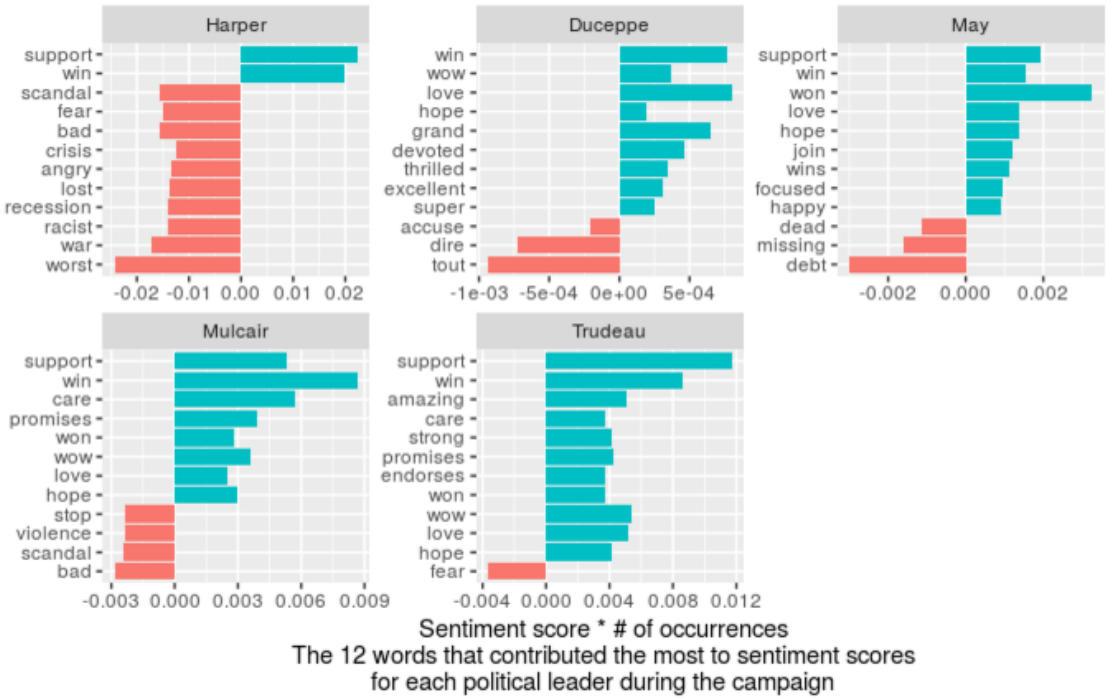

Figure 5.5 Most impactful words per candidate

5.7.2 Unsupervised Learning with LDA

After performing an LDA on the complete dataset of tweets, candidates are associated differently to the 10 topics considered. In fact, Gilles Duceppe is more associated to Topic #2; Stephen Harper to Topics #1, #3, #6 and #8; Elizabeth May to Topics #2, #3 and #4; Thomas Mulcair to Topics #4, #5 and #9; Justin Trudeau to Topics #5 and #10 (Fig. 5.6).

After removing the most common words of each topics (such as the official hashtags used during the campaign), it is possible to characterize the association of words to each politician. Table 5.2 summarizes the associated words describing each topic.

5.8 Discussion and Conclusion

In this study, we used a two-step analysis of messages published on Twitter during the 2015 General election of Canada. By focusing on the content of messages related to each politician,

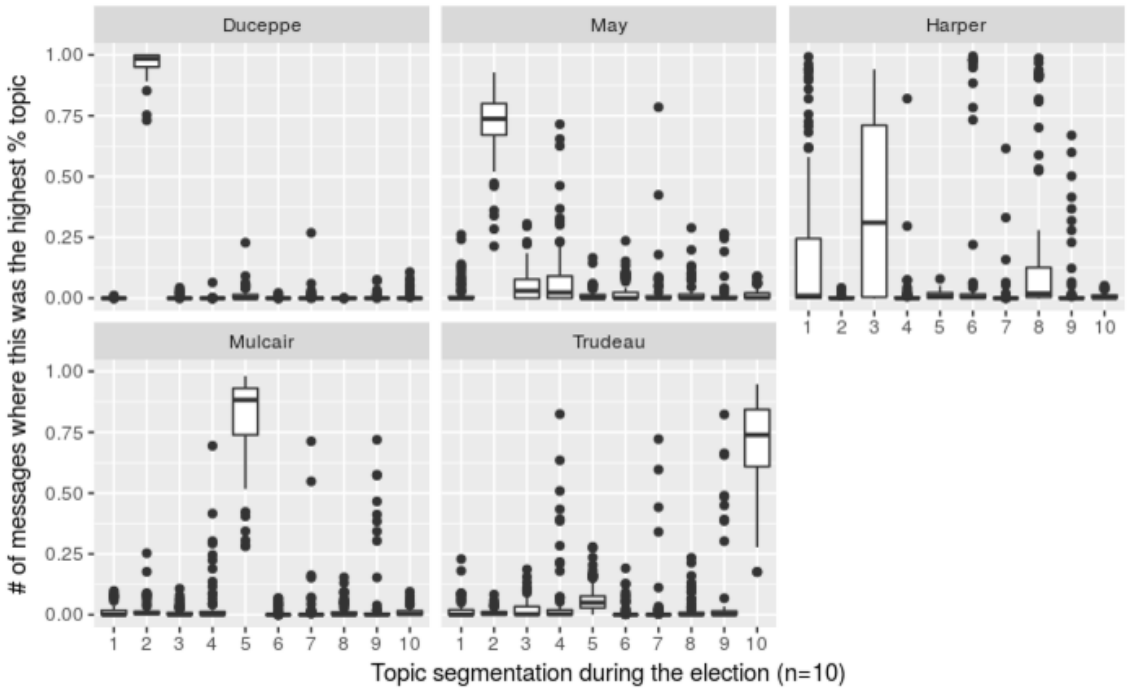

Figure 5.6 Results of the LDA analysis

we aimed to provide information regarding the evolution of the electoral campaign. On one hand, the framed strategy based on a sentiment analysis of all messages regarding each political leader have illustrated fluctuations on the perception of politicians. On the other hand, using an unsupervised methodology enables to identify topics related to each politicians without the explicit intervention of an external observer.

From our study, we can draw a few limitations that could improve further research. First, our sentiment-based scores were obtained through a unigram tokenization method. By considering bigrams or trigrams, more context may be provided regarding each political leader. Moreover, such sentiment-based scores are only as best as our predictors, namely our lexical databases. A more detailed version would increase the accuracy of our analysis, while taking into account several languages (such as French). Finally, concerning our database, we have adopted a framing strategy based on three of the most used hashtags of this election. While largely accepted in the literature, these messages did not capture every moment of the election. For example, tweets naming solely a political party may have not been taken into account. Moreover, televised debates generated an impressive participation compared to normal days (as could be seen by the five peaks in Fig. 5.3). These debates were framed by the political parties and the television networks with specific hashtags in mind.

Studying these specific moments (debates) or specific behaviors (generation of algorithmic

Table 5.2 Words associated to each topic

Number of topics	Associated words
Topic #1	Duffy trial, recession, Harper, #stopharper
Topic #2	Debate, GreenParty of Canada, Bloc Québécois
Topic #3	Refugees, campaign
Topic #4	MacLean's Debate, economy, environment
Topic #5	Mulcair, #ready4change, NDP
Topic #6	Rob Ford, #denounceharper
Topic #7	#globedebate, #faceafacetva, #oldstockcanadians
Topic #8	Harper, Mike Duffy, taxes
Topic #9	#munkdebate, #debatdeschefs, niqab, french
Topic #10	Trudeau, liberal, #realchange

propaganda) would be a great opportunity for future research in order to observe online dynamics during critical moments of a political campaign. The use of computational methodologies provide insights on how unstructured data on social media are allocated regarding public figures such as politicians during elections.

CHAPITRE 6 ARTICLE 3: NIGERIA'S 2015 PRESIDENTIAL ELECTION: A SPATIAL AND ECONOMETRIC PERSPECTIVE BASED ON A FRAMING STRATEGY

6.1 Présentation de l'article

Référence. Sanger, W. et Warin, T. 2019. Nigeria's 2015 Presidential Election : A Spatial and Econometric Perspective Based on a Framing Strategy. International Journal of Social Research Methodology. Soumis.

Le troisième article de la thèse se concentre sur l'utilisation de données issues des médias sociaux pour pallier au manque d'informations. L'élection considérée a été l'élection 2015 au Nigeria. Ce pays a été choisi puisqu'il présente plusieurs caractéristiques intéressantes au niveau démographique. La population est fortement connectée au réseau téléphonique, les usages des téléphones peuvent s'apparenter à ceux d'autres populations des pays développés et les deux candidats à l'élection ont effectué une campagne électorale en utilisant les médias sociaux pour promouvoir leurs messages politiques.

Toutefois, les instituts de sondages ne permettaient pas de suivre objectivement le cours de l'élection. Cette difficulté d'obtention d'informations fiables fut la raison principale d'étudier cette élection par le prisme des médias sociaux. Même si l'utilisation de ces données massives comporte des biais, la tenue de sondages aussi présente certains biais. Dans de tels cas, est-il alors possible de compléter les données existantes par l'utilisation de données non structurées provenant des médias sociaux ?

Pour répondre à cette question de recherche, nous avons assemblé deux bases de données. La première est constituée de plus de 550 000 messages géolocalisés publiés sur Twitter et ayant été émis depuis le Nigeria. La seconde base de données agrège des messages concernant l'élection en particulier, avec un total de plus de 1,5 millions de messages concernant les deux candidats principaux. L'objectif principal de cet article a été de recréer des mesures de suivi électoral et de développer une méthodologie économétrique pour le traitement de données issues des médias sociaux.

Cet article est coécrit avec Thierry Warin et a été soumis à l'International Journal of Social Research Methodology à l'automne 2018.

6.2 Abstract

Purpose. In countries where polling agencies do not – or barely – exist, can we use social networks as a substitute ?

Design and methods. This paper gathered two datasets of Twitter messages regarding the 2015 Nigerian election held in March 2015 (by location and by topics regarding the election). To structure these conversations, each candidate was associated to four main electoral topics with econometrics methods.

Findings. By mapping geo-located tweets, a change in the behaviour of Twitter users have been noticed a few days prior the election date. Each candidate was preferably associated to certain topics during the campaign on Twitter.

Originality and value. Evidences of the opportunity of using social networks for elections (and their limitations) were provided, as well as a methodology to structure and analyze conversations on Twitter.

Keywords. Social media, Election, Nigeria, Text analysis

6.3 Introduction

During a general election, can we use social media to get the information we would get otherwise by polling agencies in countries where polls are not used in a frequent manner ? In March 2015, Nigeria faced a presidential election. In this fast emerging country, polls are still very scarcely used during elections. It is thus difficult for a party to know whether its electoral platform works with the electorate and also to have some hints about who is leading the election at any point in time during the electoral campaign. Although imperfect, polls provide some information about the dynamics of an election. But in countries where polling agencies do not, or barely, exist, can we use social networks as a substitute ?

Indeed, it is often said that developing countries should not replicate the developed countries in all their dimensions. The argument is that developing countries can very often leapfrog. The question is thus to know whether developing countries can actually use social networks to substitute for the information polls would provide to them. Moreover, using Internet in election time provides a double positive effect [Stockemer, 2018] : it triggers for "fairer and more transparent elections" while its effect on electoral integrity is similarly strong for democracies and non-democracies.

Leapfrogging works when developed countries are in a technological or process lock-in. In this context, developing countries should go to the next technology rather than just copying and pasting the developed countries' technology. In the context of elections, if social networks are seen as a potential option for leapfrogging, this infers that we make the assumption that polls are a technological lock-in.

Social networks are more and more scrutinized and analyzed by scholars, and there is enough evidence to think that the right methodology will be found to get at least as much and as good information as the one coming from traditional polls. Twitter has attracted increasing attention in the social and information sciences as a source of data that makes it possible to gain insights into emerging social structures and content in networks, as well as community dynamics online [Pearce *et al.*, 2014].

Therefore, our research question is to know whether the information coming from Twitter can help us better understand the dynamics of an electoral campaign. More precisely, the study field considered will be the 2015 Nigerian presidential election.

This research question has one related sub-question : what topics are associated to each candidate, and how that association is evolving throughout the electoral campaign ?

In terms of methodology and empirical strategy more specifically, we use a two-stage approach. The first stage consists in framing the tweets as in [Elff, 2013]. The second stage

consists in multinomial logistic estimations to sort the content of the tweets. The justification will be presented in the methodology section.

Indeed, the goal of this paper is to offer a spatial and econometric perspective based on a framing strategy. We could have used clustering techniques or semantic analysis (LDA, etc.) of the content of the tweets, but we have decided to explore another avenue and propose the use of multinomial logistic estimations to sort out the content. Although there are limitations to this avenue, the main benefits are two fold : (1) the algorithms used in this paper are very efficient in terms of computing speed, and (2) multinomial logistic estimations are well documented in terms of their assumptions and constraints and are very robust. They are also very appropriate for our dataset, which is the result of the framing stage.

As such, this paper makes two contributions : one is methodological and the other one is substantive. The latter is particularly interesting since people tweeting about Nigeria's elections represent a natural experiment. It is also a very interesting case study, knowing that the incumbent did not win the elections.

This paper is organized as follow : in the second section, the literature review regarding the use of Twitter in electoral studies will present works of scholars, as well as their limitations and what studies have emerged in this field in Nigeria during the last two elections. After that, a spatial analysis of the 2015 Nigerian election will be presented. Section four details the dataset that we aggregated, explains the methodology used in order to tackle the sub-question and the results we obtained.

6.4 Literature Review

With the advent of new technologies and the social media revolution, we have access nowadays to an overwhelming amount of texts, while also having access to the technology to deal with it. The first part of this literature review will shed light on works regarding Twitter and elections, in the second part, the limitations of such studies will be provided and finally what has been done in Nigeria during the last two election.

6.4.1 Twitter and Elections

Although the use of Twitter might be a little bit more than a decade old, politics and social media have generated a prolific research. This is partly due to the open characteristics of the platform, since accessing messages is possible through Twitter's APIs. Moreover, there are several elections held worldwide, offering a particularly frequent and fresh source of data. In what follow, we will present the main approaches developed by scholars regarding this field

of study.

Text analysis has always been part of the toolkit of social scientists. For instance, in Political Science, Members of Congress invest substantial resources to communicate with constituents, issuing thousands of statements, press releases, and speeches during each legislative term. Those texts are an incredible source of information, which - well analyzed - could explain the dynamics of the information transmission between constituents and legislators [Grimmer, 2010].

President Obama's 2008 campaign was coined as the perfect example of an early adoption of social media at a massive scale. From "Yes We Can" slogans to an extensive Facebook presence from the Democratic candidate, the elected President has since popularized social media as a vector of expression, but moreover as a way of reaching and interacting with the citizens. The success in the use of social media has been greatly attributed to the fact that it was introduced early in the process of the election. By connecting millions of users, [Cogburn et Espinoza-Vasquez, 2011] described the U.S. as a "networked nation", which facilitated passing or approving policy issues.

A central question in research regarding social networks and Twitter in particular is the reliability of social media to serve as predictive tools for elections outcomes. A metric used to assess the accuracy of such predictions is the mean absolute error (MAE), defined as the average of the errors for each forecast concerning candidates or parties. For example, [Tumasjan *et al.*, 2010] obtained a MAE of 1.65% (Germany, with a method which has been contested since [Jungherr *et al.*, 2012]) and [Livne *et al.*, 2011] predicted 88% of the election outcomes (U.S.), offering some promising perspectives on the use of Twitter for politics. For the Irish election of 2011, [Bermingham et Smeaton, 2011] obtained a MAE of 5.85% after collecting and analyzing 32'000 messages, which is slightly higher than polling errors. [Prasetyo, 2014] surveyed 21 different papers covering different elections and found a MAE ranging from 0.1% up to 39.6% in the literature depending on the method used for predicting the outcome of election. In the Indonesian context, while testing how effective tweet-based election prediction performed compared to offline polls, the author specified that with a diversified country-level database of tweets, tweet-based prediction can outperform offline predictions (200'000 tweets from 70'000 different users).

There are multiple ways of extracting information on Twitter : amongst them, (1) the number of followers could be considered as a simplistic definition of reputation and attention ; (2) the number of mentions could be also interpreted as essential in order to analyze discussions generated on the social media ; (3) the sentiment associated to these messages may retains value and finally (4) a graphical approach could be taken in order to visualize the network

of users and how each user interacts with others.

In 2010 the Australian Federal Election was held and a paper written by [Bruns et Burgess, 2011] focused on the evolution of topics during the campaign preceding the elections. They extracted several metrics regarding their tweet sample (415'000 tweets with the hashtag #ausvotes). For instance, 35% of the messages sent prior the elections were in fact replies (retweets, with the mention "RT"), while 20% were messages addressed to specific users (with the mention "@"). A research from [Song *et al.*, 2014] reveals some topic patterns in the 2012 Korean presidential election. After performing text analysis, they were able to associate presidential candidates to their most frequent co-occurring terms through time before the election date. During the latest U.S. election in 2016, [Yaqub *et al.*, 2017] explored how Twitter could serves as a mirror of the election. By studying user behaviors on the platform, they found that little original content emerged from users while topics and sentiment associated to tweets could be correlated to public opinion and events occurring during an election.

6.4.2 Twitter and Influence

Scholars also mapped the network of users retweeting and addressing each other while identifying key characteristics of the network [Burgess et Bruns, 2012]. This graphical approach has been used in finance in order to differentiate influential, talkative and critical users regarding stock listed on the S&P500 [Marcellis-Warin *et al.*, 2017]. The polarization of Twitter users has been studied also by [Conover *et al.*, 2011]. Using 250,000 tweets about the 2010 U.S. midterm elections, they performed cluster analysis and shed light on highly segregated structures between left and right users. Users preferentially retweet other users with whom they agree politically, while the networks based on each mention appeared to form a bridge between users of different ideologies : the retweet network is highly polarized, while the mention network is not. Clustering effects have been also noticed in networks of Twitter users debating about politics. During the 2011 Canadian election, supporters of the same party seemed to interact between each other. However, connections between different party sympathizers were also revealed, since similar parties tended to have positive exchanges and hostile exchanges between opposing parties [Gruzd et Roy, 2014].

Political debates are some of the key moments in an electoral campaign. Assessing the reaction of television viewers is of great value for political parties since they have the ability to react and frame their messages. By its real-time nature, Twitter could be helpful in doing so. In 2008, the debate between Canadian party leaders was analyzed through comments on Twitter [Elmer, 2013]. In Norway, the 2011 election presented two television debates. Scholars found

that Twitter discussions reflected topics opposing candidates on television. However, the social media served as a channel for criticizing the debates but also for supporting candidates [Karlsen, 2011].

In a study published in 2014, [Pearce *et al.*, 2014] analyzed the public debate regarding climate change through messages published on Twitter. One of the approaches was to focus on the interaction between online users in order to reveal communities of individuals. They also highlighted the distribution of messages per user as being highly skewed, i.e. a few users generating the most content. Such consideration was also found in [Bruns et Stieglitz, 2013] or in [Cha *et al.*, 2012], focusing on the active role of individuals. These power-law distributions are also present in financial conversations [Marcellis-Warin *et al.*, 2017].

6.4.3 Methodological Limitations

But the question of using Twitter as a predictive tool is not settled. The previous promising results have been contrasted by other studies, emphasizing the inability to predict election outcomes by using social networks. In fact, by observing the number of followers, [Cameron *et al.*, 2013] expressed concerns and found that Facebook metrics (number of friends) may be more useful in predicting election outcomes. However, the predictive power of their models is still low (16.7% for Facebook, 5.4% for Twitter). [Choy *et al.*, 2012] use weighting techniques with online sentiment to predict the vote percentage a candidate during the 2011 Singaporean election would have received. They were able to mimic the outcomes of election by age group, but the MAE obtained was as high as 6.06%, with predictions limited to only the 2 best contenders. During the 2010 US Senate election, [Chung et Mustafaraj, 2011] applied [Tumasjan *et al.*, 2010]'s methodology and could not replicate their results (the volume of tweets concerning the two candidates).

[Gayo-Avello, 2012a] illustrated this pessimistic perception of Twitter use for electoral outcomes. With more than 234'000 messages about the 2010 US Senate election, their MAE peaked at 17.1% using the number of messages, and 7.6% using an approach based on sentiment analysis. [Gayo-Avello, 2012b] presents concerns of the potential of predicting electoral outcomes with Twitter since there is a lack of balanced literature review in the field - most of mentioned contributions validate the predictive power of the social media. Also, the presence of rumours or propaganda is ignored in most of the studies. Another weakness is that the electoral demography is not respected and that a vast majority still remains silent on Twitter (or do not participate). A survey of electoral prediction using tweets [Gayo-Avello, 2013] presents mitigated results depending on the method used by the authors.

In a paper published in 2015, [Barberá et Rivero, 2015] studied the representativeness of

Twitter users during two elections (US – 2012 and Spain – 2011). While providing a demonstration on the benefit of using social media data in electoral context, they also emphasized that participation on Twitter "is not homogeneously distributed among users". In fact, a majority of participants are men and the geographic distribution of users seems to replicate the population distribution, even though urban areas are more expressed. On the discussion content, major events in the campaign (such as televised debates) drive participation. Finally, political messages are "highly polarized", i.e. well-decided users are more proactive in the online debate. Such results help structure the representativeness of the sample of data extracted from Twitter.

6.4.4 Election Studies and Social Media in Nigeria

In the Nigerian context, a few studies have been published concerning the use of social media for electoral purposes (especially Twitter). Empowerment of the electorate is linked to the use of Twitter [Ifukor, 2010]. Indeed, in a short study where the author screened the content of 923 tweets, Ifukor showed that in the 2007 Election Twitter served as a mobilization tool and enables users to act as watchdogs regarding vote counting. With the same scope, social media can overcome traditional media scarcity [Smyth et Best, 2013]. It creates a perception of enhanced transparency for population and helps defusing tension around the credibility and the acceptability of election results [Odeyemi et Mosunmola, 2015]. In studying how social media were used in the Nigerian's political context, [Olalekan, 2015] highlighted 6 specific roles : documenting political activities ; coordinating events ; serving as a forum for citizens ; creating a platform for such citizen to communicate with the government ; checking how the public think about political decisions and being an alternative news provider.

[Fink *et al.*, 2012] developed a precise study of the use of Twitter in Nigeria using geo-located messages. 107 millions tweets were gathered and analyzed during a year, shedding light on behaviours during the 2011 Election. 246,000 users were concerned by this study, where the authors streamed messages published by 45 different cities in the country. They found that the mentions of candidates matched regional trends but that sentiment analysis provided worse predictive results. Finally, geo-located users of Twitter match the distribution of the population in the country. In a subsequent study using the database from the same research project, [Fink *et al.*, 2016] explored how 42 hashtags were adopted by the Nigerian population. They described the characteristics for a local social movement to transform itself in a widespread phenomenon and found that "hashtags associated with participation in a costly and high-risk protest movement are more likely to emerge from close-knit social communities".

To summarize, scholars have used social media (and Twitter) as a promising gateway to access new level of information in electoral context. Hence, we gathered two sets of data in order to tackle our sub-question.

6.5 Spatial Analysis of the 2015 Nigerian Election

First, let us have a look at some descriptive statistics about the general population and social media in Nigeria. Table 6.1 presents the distribution of population using the Internet versus the distribution of the general population. While Internet users are a little bit younger and are twice likely to hold a post-secondary education degree, the distribution across the southwest zone and the north zone of the country seems to be kept.

Table 6.1 Descriptive statistics of the general population and Internet users in Nigeria

Category	General Population	Internet Users
Mean age	30.6	28.2
% with postsecondary education	31.0%	66.8%
% urban	76.9%	92.1%
% male	50.7%	65.2%
% in Southwest zone	20.5%	21.5%
% in North zone	49.8%	43.6%

Finally, it is interesting to note that in comparison with the United States' population, even though a lower percentage of the population has access to Internet, Nigerian users are more likely than U.S. users to (1) use social media and (2) share politically related subjects via social media (49% in Nigeria vs. 38% in the U.S., Table 6.2).

Table 6.2 Comparison of the use of Internet and social media between the Nigerian and the U.S. populations. Source: Pew Research Center, 2015

Category	Nigeria	U.S.
% of the population having access to Internet	39%	84%
% of the Internet users using social media	82%	74%
% of users promoting political subjects via social media	49%	38%

Three important dates have to be taking into account when studying the 2015 presidential election of Nigeria. The initial date of the election was set on February 14 th 2015. However, two events led to the postponing of the election : first, Boko Haram's activity on the Northeastern part of the country focused the resources of the executive power during that time.

Also, there was an ineffective distribution of the voting cards to electors in that region. On February 8 th , due to these two events, it was decided that the elections would not take place on the 14 th of that month, but instead six weeks later, on March 28 th 2015. These events, along with concerns about biases from the police and the army, as well as campaigns using arguments towards or against religious and ethnic specificities of the population, could have raised questions on the electoral process [Onapajo, 2015]. Although the Election Day was followed by positive outcomes, the author analyzed the quality of the 2015 election and stressed critical areas of concerns at the pre-election phase of the electoral process.

Interestingly, the use of Twitter has changed during this 6-week postponing by supporters of one of the candidates during that time. In fact, when the initial date of the election was still settled, people were tweeting using the hashtag #febuhari, in reference to Muhammadu Buhari, leader of the All Progressive Congress (APC). A month later, this hashtag was left in favour to another one, #marchforbuhari, in reference to both the new election date but also to a general movement towards change.

During this election, many parties battled for power, but two major parties emerged. One the one hand, the incumbent candidate, Goodluck Jonathan, led the People's Democratic Party (PDP). Former Vice-President during Yar' Adua's presidency (2007-2010), he replaced him due to medical treatments. Jonathan is of Christian confession, and was elected president in 2011 after winning against Muhammadu Buhari with an important advance in the voting results (59% of the electorate).

His main opponent is Muhammadu Buhari, leader of the APC. Of Muslim confession, he is a former general of the Nigerian army and took control of the country after a military coup from 1983 to 1985. After a detention of 3 years due to another military coup that overthrown him, he tried to be elected president several times : in 2006, he lost against Yar' Adua (70% of the vote for Yar' Adua versus 18% of the vote for Buhari) and in 2011 he lost against Goodluck Jonathan.

As for predicting the outcome of the 2015 Nigerian Election, a few polls have been conducted by websites, radios or independent agencies. However, methodological problems have to be noticed since a general lack of information regarding these polls existed, especially concerning the number of respondents or how such polls were conducted. Table 6.3 presents some of the polls found before the elections of 2015.

The incumbent (Goodluck Jonathan) was defeated by Muhammadu Buhari after 6 more weeks of electoral campaign. APC won with 53.96% of the voting share, while PDP gained 44.96% of the voting share. Buhari obtained most of the electoral states. This diminishes the impact of such polls, since this results had not been anticipated, nor the magnitude of

Table 6.3 Polls conducted before the 2015 election

Poll source	Date	Respondents	Undecided	Buhari	Jonathan
Sahara Reporters	October 15th, 2014	15 435	/	79%	21%
Buildup Nigeria	October 16th, 2014	26 595	2.29%	48.41%	49.3%
Afrobarometer	December 5th-27th, 2014	2 400	11%	42%	42%
Nigerian FM	December 22th, 2014			64%	36%
NigerianEye	January 20th, 2015	7 043	/	72%	25%
WorldStage Newsonline	March 27th, 2015	18 866	/	35.53%	64.48%

which Buhari won the election. Hence, we are interested in establishing if Twitter could help understand the dynamics of electoral campaign when confronting to a lack of reliable polling agencies.

To provide an answer to this research question, we collected and analyzed messages related to the 2015 Nigerian election during the last 27 days of the electoral campaign. In Figure 6.1, we present the number of messages collected by day during the time of the study.

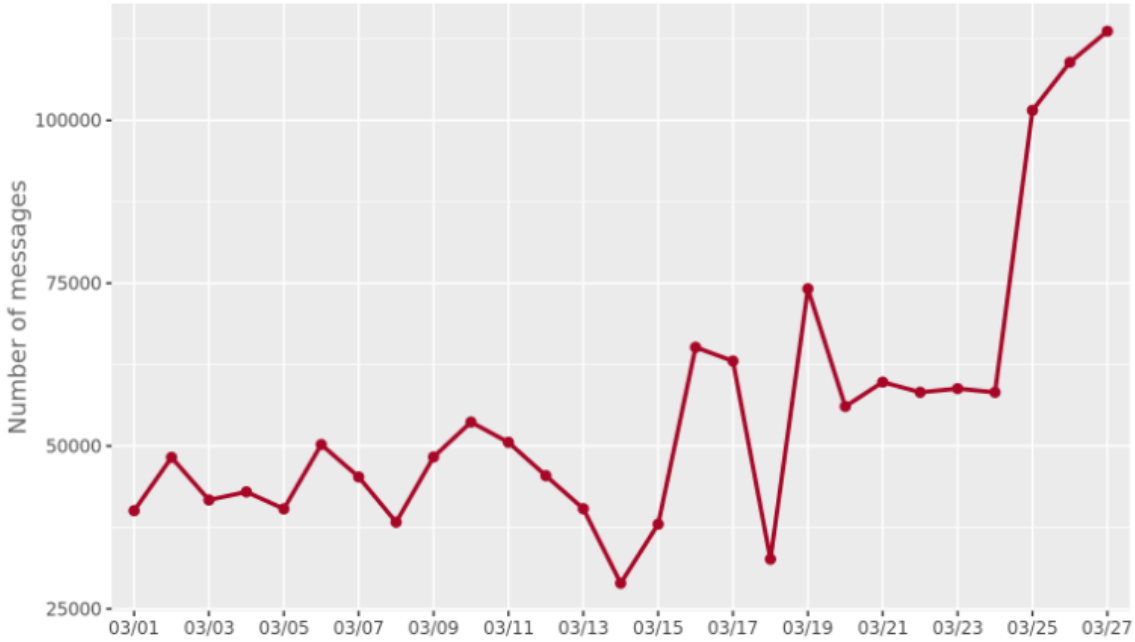

Figure 6.1 Data collection through the election period

This dataset is composed of messages emitted from Nigeria. Using the REST API of Twitter, we gathered messages during the last 27 days of the electoral campaign. Each of these messages had to be sent within a particular geographic area, which is delimited by the following

coordinates couples :

- Longitude : 2.56 ; latitude : 15.02
- Longitude : 2.56 ; latitude : 13.93
- Longitude : 4.24 ; latitude : 15.02
- Longitude : 4.24 ; latitude : 13.93

Figure 6.2 represents the area of data collection, centered on Nigeria and its surrounding countries.

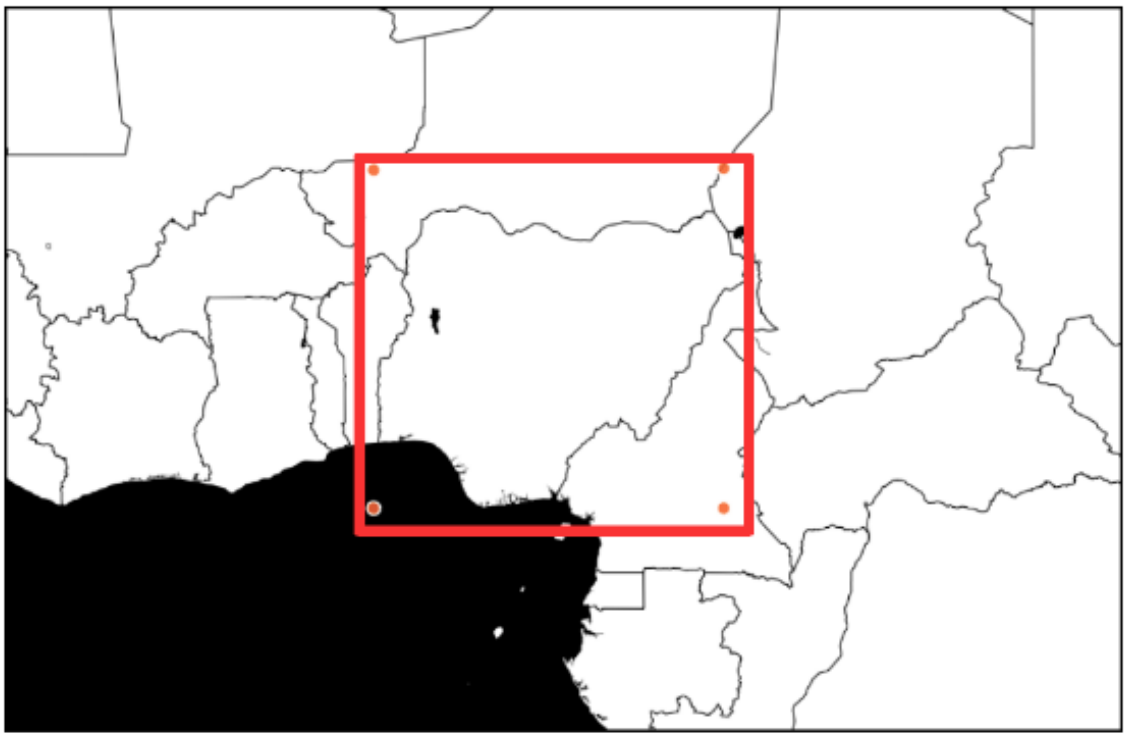

Figure 6.2 Area of data collection, centered on Nigeria and its surrounding countries

A total of 555,811 messages were collected during that period of time. Messages emitted from nearby countries (for example Cameroon) were discarded from the dataset. We use a two levels mapping approach, i.e. for each message mentioning either Buhari or Jonathan, we position them on a country level and on a city level.

We found that on a country level, each candidate was mentioned across Nigeria. There was no digital division between North and South since both candidates appeared in messages emitted across the country. However, one could found not surprisingly that cities generated most of the emitted messages.

On a city level, each city presents patterns where Buhari was more mentioned than its

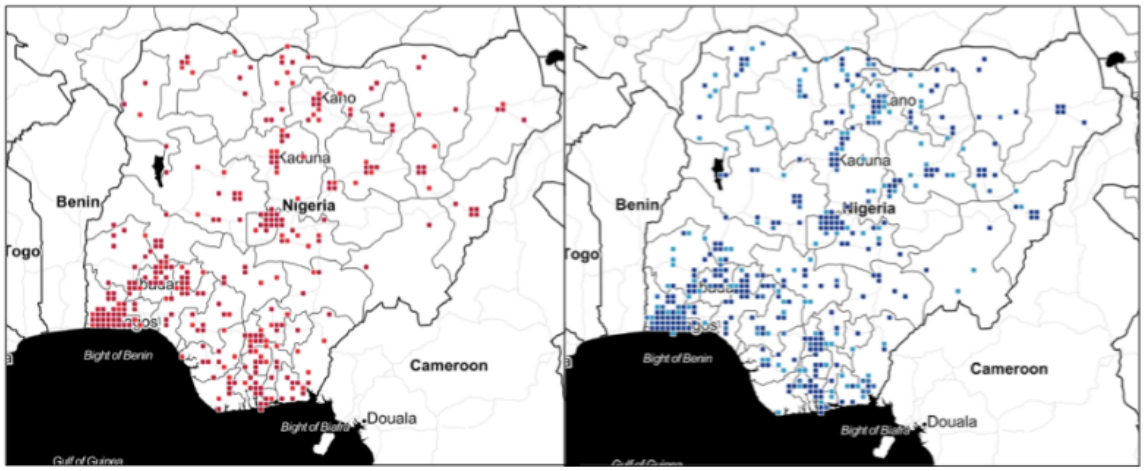

Figure 6.3 Geo-located messages for Jonathan (left) and Buhari (right) in Nigeria

competitor. In fact, by using only the number of time a candidate appeared in the geo-located messages, a clear advantage could be attributed to Buhari : he was mentioned by people spread across each city, hence implying more people mentioning him. Both the number of messages and the area of the concerned population were in favour of Buhari.

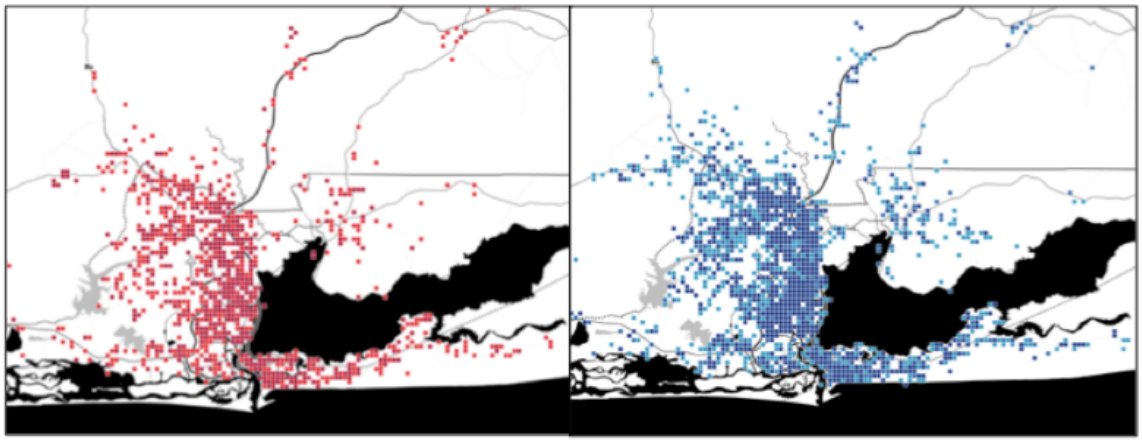

Figure 6.4 Geo-located messages for Jonathan (left) and Buhari (right) in Lagos

It is interesting to note that this relation is confirmed in Lagos, the economic capital of the country, as well as in Abudja (capital of the country) or in a northern city, Kano.

Using the same dataset, we computed three indicators, namely the number of messages concerning each candidate, the number of different users mentioning them and the average number of messages emitted by user per day.

While during the last 27 days of the electoral campaign each candidate generated almost

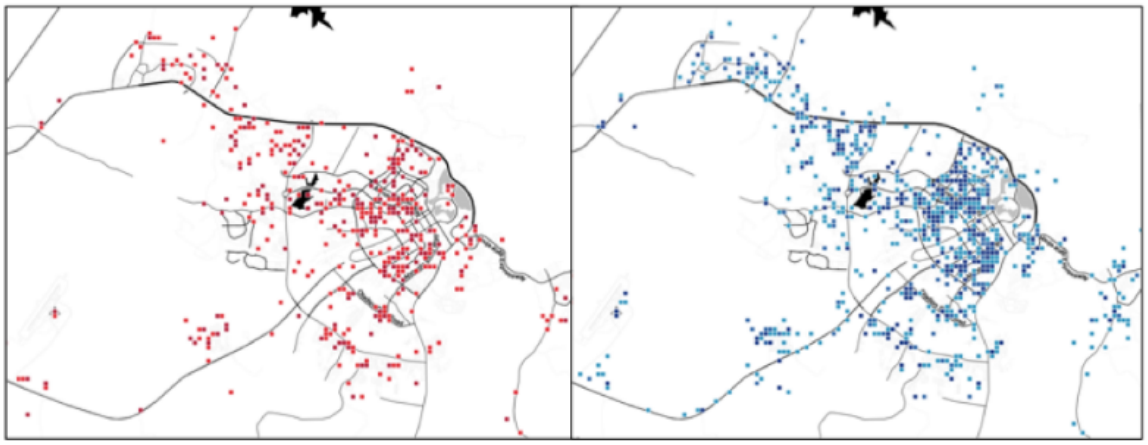

Figure 6.5 Geo-located messages for Jonathan (left) and Buhari (right) in Abudja

Figure 6.6 Geo-located messages for Jonathan (left) and Buhari (right) in Kano

the same number of messages or the same number of different users mentioning him, it is interesting to mention a shift 5 days prior the Election Day. In fact, Buhari seemed to emulate more messages about himself than its opponent, as well as implying more people in the discussions concerning himself in Nigeria (respectively Figures 6.7 and 6.8).

When plotting the average number of messages by user on a daily basis, such shift is noticed around that date, between March 21st and March 22nd (Figure 6.9). More investigation may be useful in order to fully understand why such behaviours happened, but can reveal a trend towards Buhari, which subsequently was the winner of the election.

In conclusion, by using geo-located message, it is possible to (1) map where tweets where emitted from, but also to (2) notice how a particular candidate (Buhari) generated more

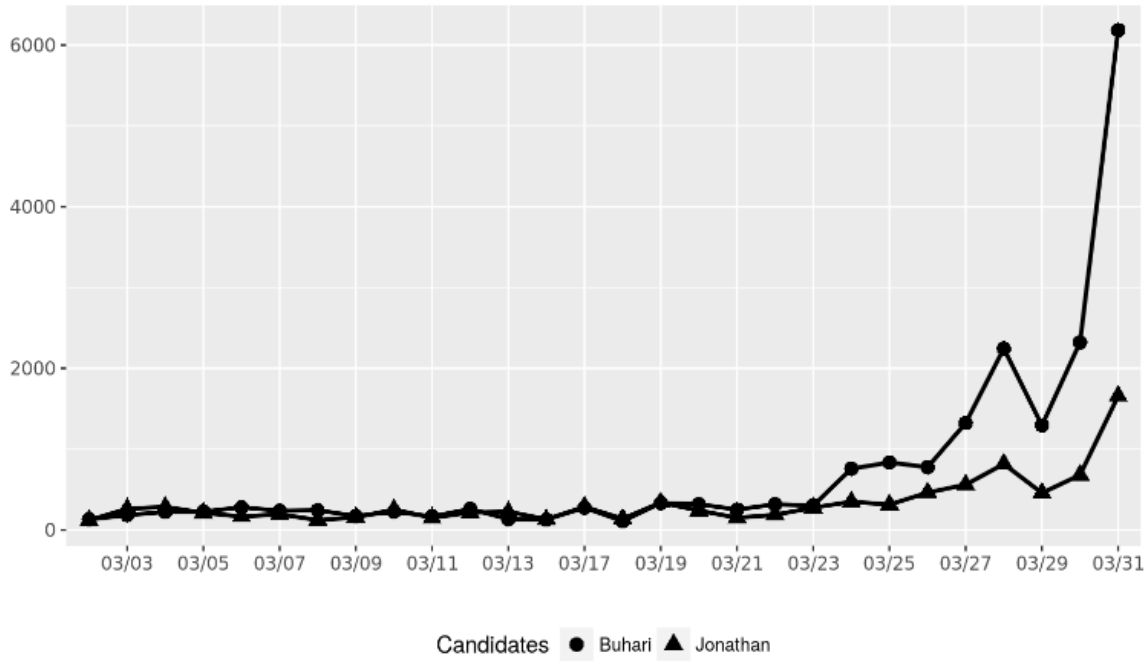

Figure 6.7 Number of geo-located messages per candidate

attention then it's opponent. A correlation with the number of messages sent daily and the results of the election is also noticed.

6.6 Empirical Strategy

Let us present our pre-processing strategy and some descriptive statistics first.

6.6.1 Pre-processing and Descriptive Statistics

The dataset is composed of messages published during the last 27 days of the election campaign. Using the REST API from Twitter, several keywords were selected to collect messages related to the 2015 Nigerian Election. A first set of keywords were considered, such as "Nigeria", "Boko Haram", "Buhari", "Goodluck Jonathan" and tweets published by influent members of each political parties (and all derivatives, including the previous keywords with hashtags).

We use a framing strategy to assemble our datasets. It has some limitations, which we need to take into consideration to have an efficient unbiased collection of data. For instance, in Political Science, framing has been used in many instances. [Elff, 2013] uses this technique for the reconstruction of parties' political positions based on coded political texts. The author

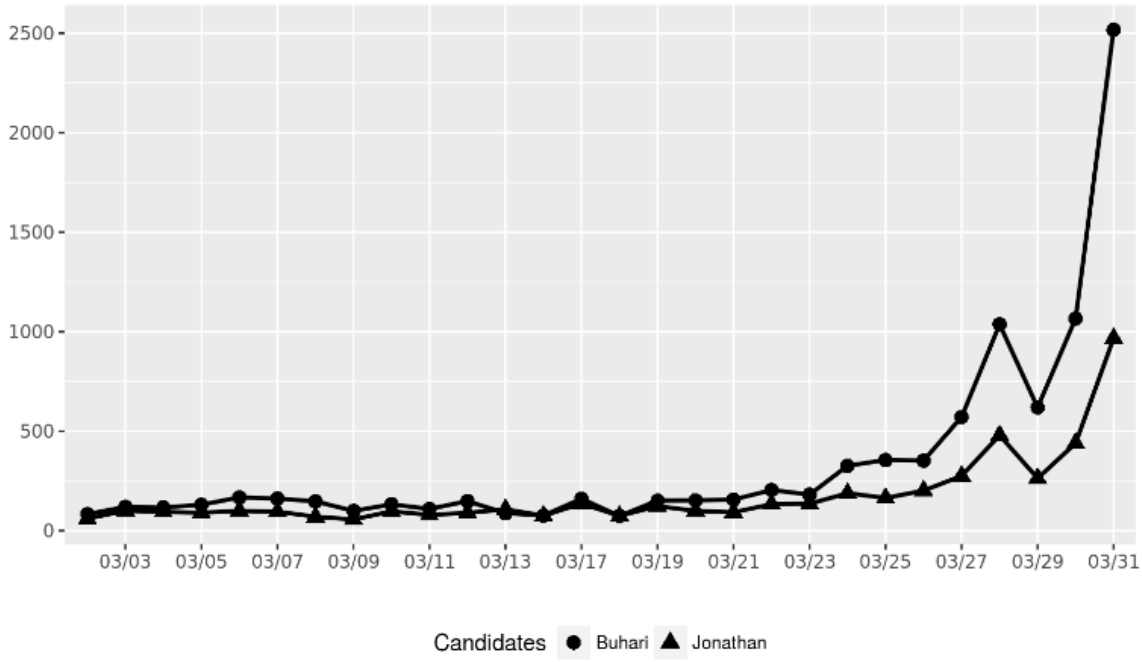

Figure 6.8 Number of unique users per candidate

concludes that it can only be as good as the categories used in coding them. Indeed, the coding categories have been occasionally criticized as not very well capturing what is really going on in parties' electoral platforms [Laver et Garry, 2000]. [Chong et Druckman, 2007] note that the analysis of "frames in communication" is an endeavor that "has become a virtual cottage industry." In such analyses, "an initial set of frames for an issue is identified inductively to create a coding scheme" [Chong et Druckman, 2007]. These then serve as the basis for a content analysis of relevant texts, like newspaper articles, to evaluate variations in frame across source and across time [Monroe *et al.*, 2008]. Improved coding schemes and procedures may thus enhance the precision and reliability of various analyses. This is what we aim to do in this paper.

This first dataset gathered more than 3.8 million messages. As a second filter, we selected within this dataset only messages that mentioned either "Buhari", "Jonathan" or the two hashtags used during the electoral campaign, "#nigeria2015" and "#nigeriadecides". This subsequent dataset gathered 1.541 million messages in total. That way, we first selected messages sent worldwide concerning Nigeria and its election. However, by selecting messages with the second list of keywords, we filtered and cleaned the dataset to only obtain messages regarding the election in particular, since the two hashtags were used during the election campaign, as well as publicized on official documentations and communications.

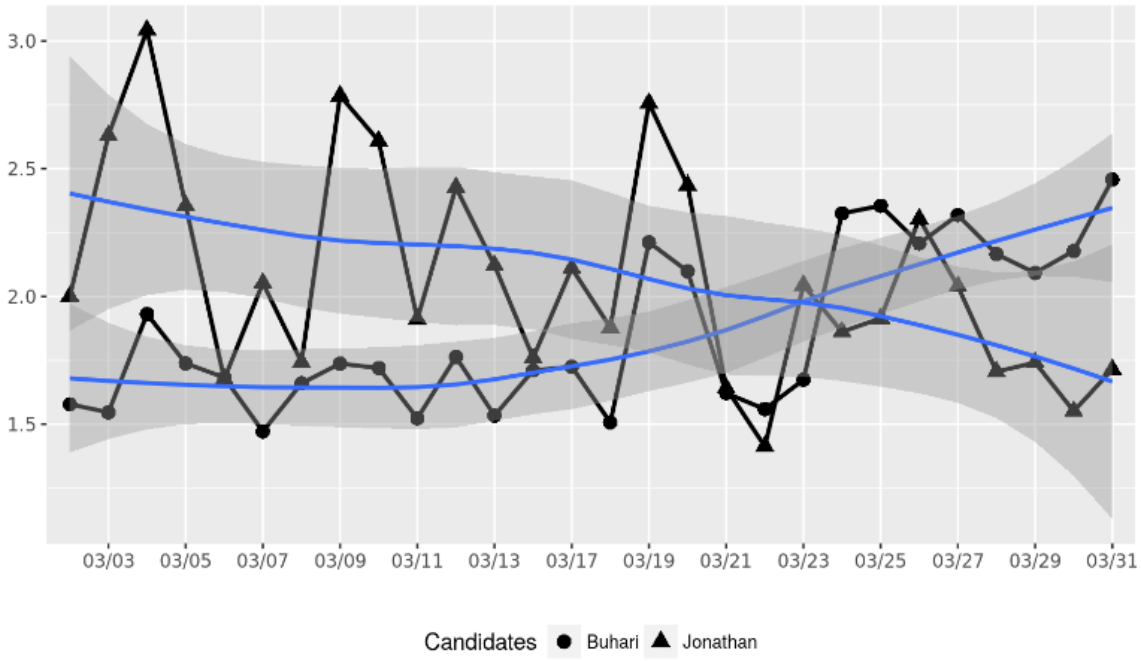

Figure 6.9 Average number of messages per user for each candidate

We structured this dataset in a cross-section time-series format. 4 main topics were selected to represent the main topics of the campaign. Each message was considered as a message related either to (1) social, (2) integrity, (3) economy or (4) geopolitics. For each main topic, we selected 5 keywords, for a total of 20 subtopics. Using regular expressions, we were able to identify a message as a message that was mentioning a particular subtopic, hence a particular main topic. For social, we were searching for messages related to “education”, “school”, “family”, “health” and “doctors”. For integrity, we were searching for messages related to “corruption”, “integrity”, “ethics”, “fraud” and “bribe”. For economy, we were searching for messages related to “economy”, “finance”, “employment”, “jobs”, and “tax”. Finally, for geopolitics, we were searching for messages related to “security”, “terror”, “war”, “Boko Haram” and “army”.

Table 6.4 gives an overview if some descriptive statistics related to each subtopic considered.

Table 6.4 Descriptive statistics of each of the main topics and each of the 20 subtopics

Keyword	Min	Max	Average	Std. Dev.	Q1	Q3
Topic 1: Social	**862**	**3700**	**1752.48**	**646.52**	**1407**	**2007**
Doctor	3	255	81.48	68.90	24	120
Education	72	1702	415.96	375.06	244	390
Family	19	560	122.74	113.47	49	142
Health	49	778	338.52	218.43	147	441
School	110	1795	793.78	398.27	522	955
Topic 2: Integrity	**716**	**5017**	**2432.30**	**1160.47**	**1602**	**3541**
Bribe	8	339	94.63	77.56	31	125
Corruption	537	4667	2163.70	1048.82	1451	3003
Ethics	0	24	2.63	4.87	0	3
Fraud	8	308	71.81	73.06	27	91
Integrity	15	379	99.52	93.31	38	136
Topic 3: Economy	**297**	**2748**	**1552.81**	**660.21**	**1237**	**2080**
Economy	147	2433	1191.37	635.84	843	1588
Employment	10	574	112.74	149.22	30	106
Finance	7	217	62.59	57.46	22	74
Jobs	11	784	142.96	177.37	41	150
Tax	4	212	43.15	51.40	15	44
Topic 4: Geopolitics	**1276**	**11153**	**4736.56**	**2925.47**	**2146**	**6348**
Army	33	605	142	140.66	63	132
Boko Haram	303	4293	1241.89	809.81	757	1472
Security	45	1111	307.37	236.03	143	403
Terror	32	1016	370.11	292.86	119	588
War	372	7580	2675.19	2116.20	954	4098

We were able to structure the dataset in a cross-sectional time-series dataset : for 27 days, we compute the number of times each of the 20 subtopics was mentioned. On Figure 6.10, the evolution of each main topic could be observed, as percentage of the whole conversation for each day. For example, on March 9th, 10.5% of the messages emitted mentioning either "Buhari", "Jonathan", "#nigeriandecides" or "#nigeria2015" were associated to the main topic related to geopolitics.

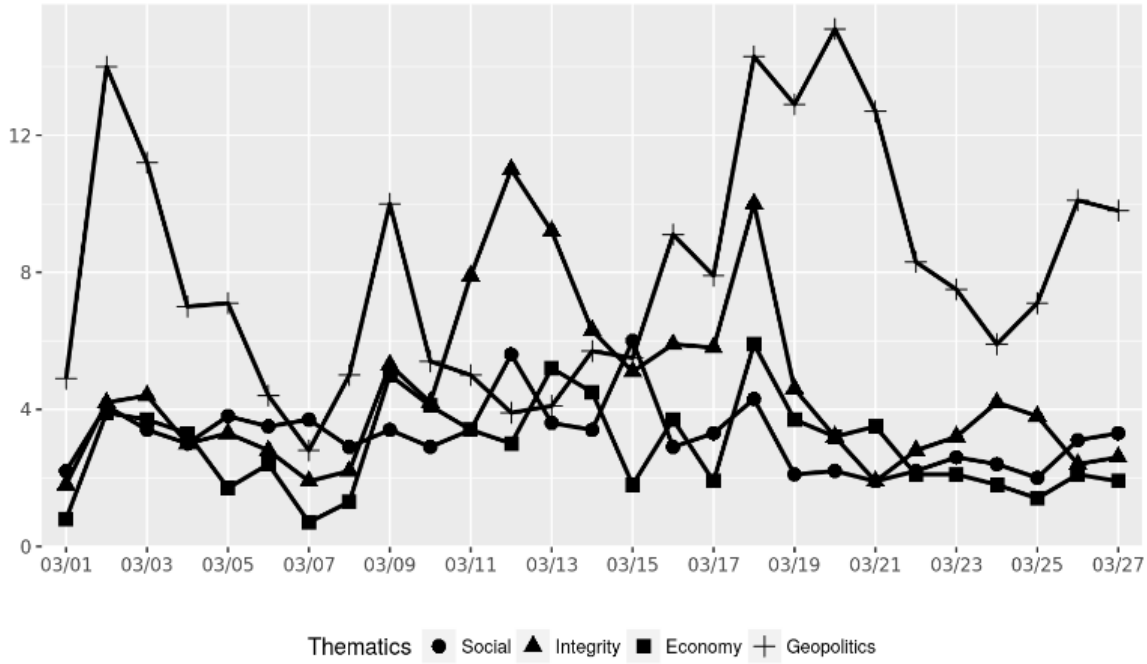

Figure 6.10 Evolution of each main topic in terms of share of conversation (%)

Within these messages, for each day and each subtopic, we compared the number of time each candidate (Buhari or Jonathan) was mentioned, and the number of times each of the political party (APC or PDP) was mentioned. In terms of methodology, we used the logistic estimators, as explained in the next section.

6.6.2 Methodology

The econometric models used in this paper were adapted from a previous article about the Province of Quebec's elections [Sanger et Warin, 2018c]. Messages about elections in general will be considered in order to assess what Twitter users are saying on the presidential candidates. This dataset will be later referred as the "topic dataset".

Based on a framing protocol developed by [Elff, 2013], we obtained for each day the main topic of the conversation (one of the four main topics). Further on, for each topic and for

each day, we observed which candidate was mentioned the most (either Jonathan or Buhari). Finally, we segmented our dataset in four periods of time (from March 1^{st} to 6^{th} ; from March 7^{th} to 13^{th} ; from March 14^{th} to 20^{th} and from March 21^{st} to March 27^{th}).

More specifically, each tweet $N \in \{1, ..., n\}$ is redefined into mutually exclusive categories representing the four main topics of the campaign $i \in \{1, ..., 4\}$ for each day $t \in \{1, ..., T\}$.

In this context, the question is to select a logistic estimation that will extract high quality information. We have to decide first between a discrete modelisation with only two outcomes or a discrete modelisation with more than two outcomes. The choice of two outcomes infers that we create four variables capturing the 4 categories of our initial dependent variable.

Secondly, if we decide to keep our category variable with 4 categories, then another decision has to be made : choosing between a multinomial logistic estimation or an ordered logistic estimation. The issue is not trivial. Indeed, although the demarcation line is often clear, in our context, we are in a grey area.

Due to the nature of the study, it is not clear what estimation technique is best. Indeed, in a traditional setting, a ranking can be done. In this example, there is no order. Here, we collect tweets and we aggregate the number of tweets at the end of the day per category. In our context, a person can tweet about a topic and then tweet at another time of day about another topic.

In many regards, it is like looking at a permanent poll, which raises interesting statistical questions and thus requires or allows for new techniques or protocols. Indeed, even if the categories we chose (social, integrity, economy and geopolitics) have no order, in fact the persons tweeting during the day make a choice like in a poll and we can assume that if they tweet more about a topic, it is because they do believe this topic matters more to them than another one. For introduction to logistic regression, see [Hosmer Jr *et al.*, 2013, Pampel, 2000] ; for a complete but non mathematical treatment, see [Kleinbaum et Klein, 2010] ; and for a thorough discussion, see [Hosmer Jr *et al.*, 2013]. Regarding a discussion of logistic regression, see [Dupont et Dupont, 2009] or [Hilbe, 2009] and for an interpretation of logistic regression, see [Gould, 2000]. If this hypothesis is right, then the next question is to know the order. A specific logistic estimator is in fact designed for this kind of characteristics : the stereotype logistic estimator. Unlike ordered logistic models, stereotype logistic models do not impose the proportional-odds assumption. Stereotype logistic models are often used when subjects are requested to assess or judge something.

So, as consequence, here are the estimations we present in this paper :

— A set of binary logistic estimations

— A set of categorical logistic estimations with different nuances as well as robustness checks :
  — A plain-vanilla - unordered - multinomial estimation
  — A mixed-ordered estimation : a stereotype-ordered logistic estimation

The multinomial logistic estimation fits maximum likelihood models with discrete dependent variables when the dependent variable takes on more than two outcomes and the outcomes have no natural ordering [Greene, 2012, Hosmer Jr *et al.*, 2013, Long *et al.*, 2006, Scott Long, 1997, Treiman, 2014].

Finally, we will also estimate a model that is a compromise between the ordered and unordered logistic estimations. We propose the so-called stereotype logistic estimation [Anderson, 1984, Greenland, 1985], because as aforementioned there is an uncertainty about the relevance of the ordering.

6.6.3 First Estimator : the Binary Logit Estimation

We estimate the model following a two-stage approach.

During the first stage, we collect all the tweets and we rank them at the end of the day in terms of the topics. If the most mentioned topic of the day is related to "social" ("integrity", "economy", "geopolitics") then $y_t = 1$ (respectively 2, 3, 4).

We use a set of maximum likelihood based regressions, specified as followed. As for the dependent variable, we consider each topic. As for the independent variables, we consider the number of tweets related to each candidate. As control variables, we used the four different time periods.

The following equation details a little more our approach :

$$y_t = \alpha_0 + \alpha_1.x_1 + \alpha_2.x_2 + \alpha_3.x_3 + \epsilon \tag{6.1}$$

With :

— $y_t = \{1; 2; 3; 4\}$ if the most mentioned topic of the day is related to respectively social ; integrity ; economy ; geopolitics
— x_1 the number of messages related to Buhari, per day
— x_2 the number of messages related to Jonathan, per day
— $x_3 = \{1; 2; 3\}$ for respectively period $\{1; 2; 3\}$

Finally, we calculate the predicted probabilities to be the most associated to a particular topic for each candidate during each period of time. This way, we were able to understand

the political campaign of each candidate through messages published on Twitter : for example, from March 1^{st} to March 6^{th} , was one of the candidate more associated to messages related to one of the four main topics or was his campaign's reception well distributed across each topic ?

Table 6.5 Odds Ratios of the binary logistic model

Dependent variable: topic {social OR integrity OR economy OR geopolitics}								
Model: binary logit	Odds-Ratio				Odds-Ratio			
Independent variables	Social	Integrity	Economy	Geopolitics	Social	Integrity	Economy	Geopolitics
Jonathan	0.9977681 ***	0.9987912 **	1.000224	1.001205 ***	0.997743 ***	0.9987617 ***	1.00023	1.001255 ***
	(0.0008427)	(0.0004724)	(0.0002578)	(0.0003328)	(0.0008464)	(0.0004766)	(0.0002616)	(0.0003398)
Buhari	1.000308	1.000941 ***	0.9993854	0.9995662	1.000311	1.000962 ***	0.9993768	0.9995407
	(0.0004616)	(0.0003443)	(0.0005302)	(0.0003436)	(0.0004633)	(0.0003468)	(0.0005337)	(0.0003479)
PDP	1.002938 **	0.9977917	1.00008	0.9990159	1.002935 **	0.9978165	1.000048	0.9990247
	(0.0012051)	(0.0013672)	(0.0006899)	(0.0006954)	(0.0012107)	(0.001367)	(0.0007029)	(0.0007099)
APC	0.9955218 **	1.000749	0.9963285 *	1.003131 ***	0.9955399 **	1.0007	0.9963318 *	1.003154 ***
	(0.0018847)	(0.0011726)	(0.0019085)	(0.0011826)	(0.001887)	(0.0011762)	(0.0019144)	(0.0011822)
periods (ref = period 1)								
period 2					0.9802216	0.8794575	1.0178	1.120562
					(0.2876011)	(0.2576892)	(0.2970309)	(0.3336496)
period 3					1.038389	0.9719715	0.9935214	0.9162648
					(0.3066134)	(0.285127)	(0.2917344)	(0.280578)
period 4					1.119173	1.040463	1.084849	0.7965988
					(0.3317443)	(0.3037566)	(0.3189801)	(0.2474592)
Predictions								
Pr(y=1)	0.2053 **	0.2302 **	0.2268 **	0.2451 **				
Pr(y=0)	0.7947 **	0.7698 **	0.7732 **	0.7549 **				
Number of observations	540	540	540	540	540	540	540	540
LR chi2	31.94	21.30	20.18	51.42	32.19	21.68	20.30	52.83
Prob > chi2	0.0000	0.0003	0.0005	0.0000	0.0000	0.0029	0.0050	0.0000
Pseudo R2	0.0526	0.0351	0.0332	0.0847	0.0530	0.0357	0.0334	0.0870
Log likelihood	-287.68904	-293.00921	-293.57165	-277.95044	-287.56532	-292.81884	-293.51303	-277.24682

P-value: *<0.1, **<0.05, ***<0.01

When we look at the odds ratios, it is interesting to see that Twitter conversations about Mr Jonathan were more associated with tweets about economic or geopolitical issues. Mr Buhari triggered conversations more about social and integrity issues. It is also interesting to note that it does not seem to change through time.

In what follow, we will compute the predicted probabilities for each main topic regarding each candidate for each time period. In order to interpret the next results, the predicted probabilities are : when one of the candidate is ahead of his opponent in terms of number of mentions, what is the probability that the most mentioned topic about him is either about society, integrity, economy or geopolitics ?

Table 6.6 Predicted Probabilities regarding each topic

	Pr(Social)		Pr(Integrity)		Pr(Economy)		Pr(Geopolitics)	
Period - Political Leader	Margin	P-value	Margin	P-value	Margin	P-value	Margin	P-value
1 - Buhari	0.2867873	***	0.2414737	***	0.2346261	***	0.2371853	***
	(0.0443793)		(0.0406832)		(0.039403)		(0.0396489)	
1 - Jonathan	0.1267496	***	0.2208945	***	0.3310479	***	0.3210243	***
	(0.0380131)		(0.0516675)		(0.061144)		(0.0605885)	
2 - Buhari	0.2797548	***	0.2279279	***	0.2514236	***	0.2395692	***
	(0.040227)		(0.0366072)		(0.0372835)		(0.0366873)	
2 - Jonathan	0.1229649	***	0.2081863	***	0.3515776	***	0.3238931	***
	(0.0366282)		(0.0491284)		(0.0612219)		(0.0597212)	
3 - Buhari	0.2932305	***	0.2330134	***	0.2425358	***	0.2316886	***
	(0.0414524)		(0.0372642)		(0.0370359)		(0.0364637)	
3 - Jonathan	0.1302539	***	0.2129528	***	0.3407611	***	0.3143854	***
	(0.0373019)		(0.0479942)		(0.0580282)		(0.0565756)	
4 - Buhari	0.2888603	***	0.2494513	***	0.2298601	***	0.2327109	***
	(0.0415778)		(0.0384443)		(0.0364766)		(0.0367381)	
4 - Jonathan	0.1278732	***	0.228397	***	0.3251554	***	0.3156228	***
	(0.0367203)		(0.0496819)		(0.0568843)		(0.0564164)	

Regarding messages associated to the first topic (social), Jonathan's campaign did not evolved during the last 27 days. In fact, when he was the most mentioned compared to his opponent, from 12.2% to 13% of those messages concerned society messages. On the other hand, when Buhari was the most mentioned, from 27.9% to 29.3% of those messages were related to society messages.

Across each period of time, from 22.8% to 24.9% of the messages when Buhari is ahead of his opponent were concerning Integrity matters. For Jonathan, it is 20.8 to 22.8% of the messages when he is ahead of his opponent.

Economy was one of the strong topics of former President Jonathan. In fact, 32.5 to 35.2% of all messages where he is more mentioned than his opponent concerned this topic (i.e. either mentioning finance, economy, tax, jobs or employment). For Buhari, it is a less important topic of his campaign, since 23 to 25.1% of messages when he is more mentioned concerned this main topic.

Finally, geopolitics was another important topic associated to former President Jonathan. On average, 32% of all messages mentioning him more than Buhari contained a subtopic related to geopolitics, such as war against terrorism and Boko Haram. For Buhari, this proportion is about 23.2 to 24% when he is most mentioned.

In order to synthesize the table, we provided a summary of the predicted probabilities of the logit regression in Figure 6.11. As a conclusion, from messages published on Twitter, Buhari presented a well distributed campaign since each topic were represented with a similar importance within the messages where he is the most mentioned person. Jonathan, on the other hand, presented a lack of identification to topics related to society or integrity, and was most associated to geopolitics or economy.

As we can observe, it is pretty resilient through time.

6.6.4 Second Estimators : the Multinomial and Stereotype Logistic Estimations

Let us now change our dependent variable and consider a multinomial logistic estimation and a variation : the stereotype estimator.

The multinomial estimator assumes that there is no order in the different categories used for the coding of the dependent variable. But for the stereotype estimator, it relies on one hypothesis : in fact the persons tweeting during the day make a choice like in a poll and we can assume that if they tweet more about a topic, it is because they do believe this topic matters more to them than another one. But unlike the ordered logit estimator, we make the reasonable assumption that we do not know all the latent variables to make a proper

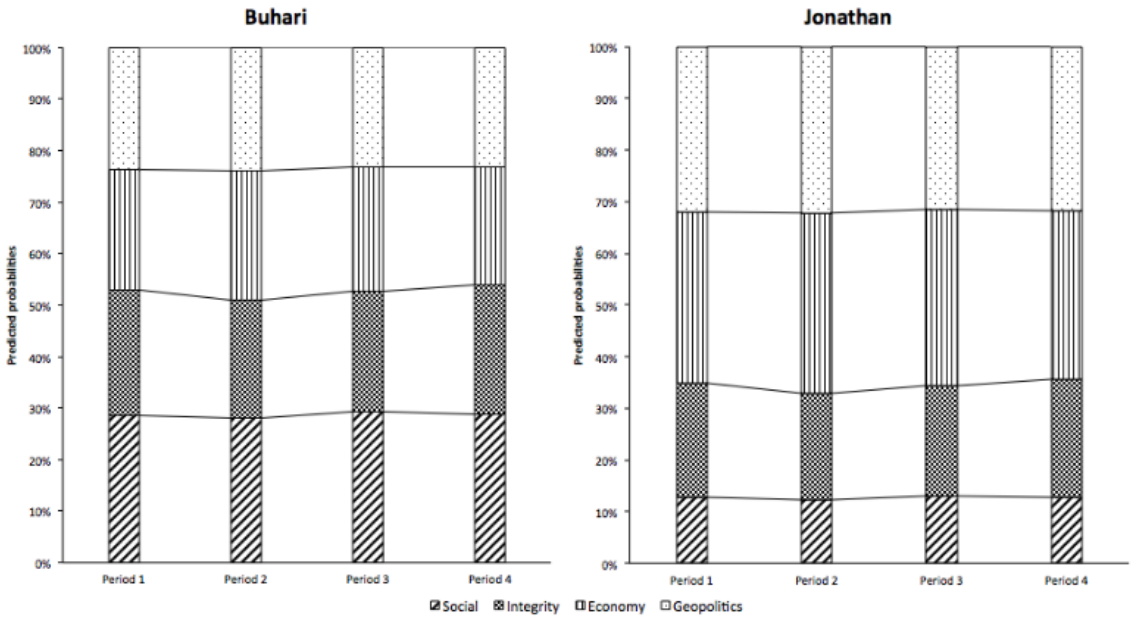

Figure 6.11 Summary of the Predicted Probabilities for each candidate for each topic

ranking.

In table 6.7, we calculate the relative risk ratios. Compared to the base outcome (social category), Mr Jonathan is less likely to be associated with the integrity category than Mr Buhari. The same is true for the economy category.

We compute the marginal effects as a robustness check. Table 6.8 shows that Mr Jonathan is more associated with the conversations about the social and geopolitics categories.

Now, let us present the results based on the stereotype logistic regressions in table 6.9. Stereotype logistic models are used in particular when categories may be indistinguishable. The stereotype logistic model should be seen as a restriction on the multinomial model.

In table 6.9, we can observe that Mr Jonathan is more associated with the conversations about the geopolitics category (coef. =0.001*), as well as the APC party (coef. =0.0036). Those are interesting results since they validate the ones we got with the plain multinomial logit estimator, although being a little less focused in terms of interpretation since they have to be interpreted vis-a-vis the base category (here, geopolitics). It is interesting anyway to be able to use a stereotype logistic estimator based on our dataset. Indeed, our framing strategy seems to allow us to perform the latter analysis and provide some robustness to the analysis of conversations on Twitter.

6.7 Conclusion

Could social media be used in order to compensate a lack of polling agencies in electoral context ? We demonstrated that it was possible to associate each candidate to topics regarding the electoral campaign.

The goal of this paper was to offer a spatial and econometric perspective based on a framing strategy. As aforementioned, we could have used clustering techniques or semantic analysis (LDA, etc.) of the content of the tweets, but we have decided to explore another avenue and propose the use of logistic estimations to sort out the content. Although there are limitations to this avenue, the main benefits are two fold : (1) the algorithms used in this paper are very efficient in terms of computing speed, and (2) logistic estimations are well documented in terms of their assumptions and constraints.

However, several limitations still exist by such methods. Although using geo-located messages enables to predict the outcome of the elections, it has to be reminded that only a fraction of Twitter users gives access to their location data when sending a message. Nevertheless, we can infer that this behaviour can be equally found in each candidate's electorate.

A second limitation is the fact that rural participants may not be well represented by such analysis. Users of social media (and Internet) are more located in urban areas. A more precise follow-up of the composition of the electorate and the use of social media through mobile phone may complement our study. Again, bias for each candidates are the same. Moreover, up-to-date statistics regarding the use of mobile phone are still lacking, which could either confirm or infirm the widespread adoption of mobile technologies.

Deeper analysis could be done regarding the 2015 Nigerian Election. One promising path is to analyze patterns of influence between users. Power-laws distributions were noticed in other context regarding the number of messages send by users on Twitter [Marcellis-Warin *et al.*, 2017], and an important role have been associated to gatekeepers on social media between groups of users. Was it the same during the 2015 Nigerian Election ?

By this study, it clearly appears that new technologies (social networks) matter, especially when lacking of efficient institutions. Hence, for countries lacking in polling agencies, using Twitter may presents opportunities to deliver new and more complex streams of information.

Table 6.7 Relative Risk Ratios and Predicted Probabilities of the multinomial logistic model

Dependent variable: topic {social; integrity; economy; geopolitics}								
Model: multinomial logit Independent variables	**Coef.**		**Relative Risk Ratios**		**Coef.**		**Relative Risk Ratios**	
Topic 1 (social)	**base outcome**							
Topic 2 (integrity)								
Jonathan	-0.0027595	***	0.9972443	***	-0.0028211	***	0.9971828	***
Buhari	0.0011461	*	1.001147	*	0.0011777	*	1.001178	*
PDP	0.0031242	**	1.003129	**	0.0031851	**	1.00319	**
APC	-0.0010757		0.9989249		-0.001078		0.9989226	
Periods (ref = Period 1)								
Period 2					-0.030186		0.970265	
Period 3					0.0424364		1.04335	
Period 4					0.0435374		1.044499	
Constant	0.0271056				0.0124693			
Topic 3 (economy)								
Jonathan	-0.001625	***	0.9983764	***	-0.0016934	***	0.998308	***
Buhari	0.0013541	**	1.001355	**	0.0014034	**	1.001404	*
PDP	-0.0014103		0.9985907		-0.0012894		0.9987115	
APC	0.004636	**	1.004647	**	0.004534	**	1.004544	**
Periods (ref = Period 1)								
Period 2					-0.1076294		0.8979603	
Period 3					-0.000384		0.9996161	
Period 4					-0.0027362		0.9972675	
Constant	-0.3239868	**			-0.294147			
Topic 4 (geopolitics)								
Jonathan	0.000212		1.000212		0.0002218		1.000222	
Buhari	0.0005398		1.00054		0.0005451		1.000545	
PDP	-0.0004569		0.9995432		-0.0003948		0.9996053	
APC	0.0055712	***	1.005587	***	0.0055733	***	1.005589	***
Periods (ref = Period 1)								
Period 2					0.097285		1.102174	
Period 3					-0.0254721		0.9748496	
Period 4					-0.2292305		0.7951452	
Constant	-0.5721883	***			0.5426475	*		

Predicted probabilities				
Pr(y=1)	0.251581	***	0.2520359	***
Pr(y=0)	0.748419	***	0.7479641	***
Topic = 1 (social)	0.251581	***	0.2520359	***
Topic = 2 (integrity)	0.2182243	***	0.2184104	***
Topic = 3 (economy)	0.2602211	***	0.2602559	***
Topic = 4 (geopolitics)	0.2699736	***	0.2692978	***

Number of observations	540	540
LR chi2	96.93	98.79
Prob > chi2	0.0000	0.0000
Pseudo R2	0.0647	0.0660
Log likelihood	-700.1345	-699.20412

P-value: *<0.1, **<0.05, ***<0.01

Table 6.8 Marginal Effects of the multinomial logistic model (without the period)

Dependent variable: topic {social, integrity, economy, geopolitics}								
Model: multinomial logit	**Marginal Effects**							
Independent variables	**Social**		**Integrity**		**Economy**		**Geopolitics**	
Jonathan	0.0002435	***	-0.000391	***	-0.000171	*	0.0003185	***
Buhari	-0.0001882	*	0.0000868		0.0001577	**	-0.000563	
PDP	-0.000482		0.00064	***	-0.0004168		-0.000175	
APC	-0.0006228	*	-0.000775	**	0.0005621	**	0.0008357	***
Number of observations	540							
LR chi2	96.93							
Prob > chi2	0.0000							
Pseudo R2	0.0647							
Log likelihood	-700.1345							

P-value: *<0.1, **<0.05, ***<0.01

Table 6.9 Coefficients of the stereotype ordered logit model, with and without constraint

Dependent variable: topic {social, integrity, economy, geopolitics}				
Model: stereotype ordered logit Independent variables	**Without constraint Coef.**		**With constraint Coef.**	
Jonathan	0.0011052	***	0.000968	**
Buhari	-0.0001663		0.0000215	
PDP	-0.0013216	**	-0.0012697	
APC	0.0036409	**	0.0055193	***
/phi1_1	1	***	1	***
/phi1_2	1.856209	***	1	***
/phi1_3	0.5181356	**	0.3122322	***
/phi1_4	0	/	0	/
/theta1	0.4423166	***	0.659083	***
/theta2	0.607536	***	0.659083	***
/theta3	0.2966307	**	0.2649479	*
/theta4	0	/	0	/

(category 4 is the base outcome)

Number of observations	540
Wald chi2	17.69
Prob > chi2	0.0014
Log likelihood	-715.05661

P-value: *<0.1, **<0.05, ***<0.01

CHAPITRE 7 ARTICLE 4: TEXT-AS-DATA ANALYSIS OF POPULIST PARTIES VERSUS GOVERNMENT PARTIES: TO BLEND OR NOT TO BLEND ?

7.1 Présentation de l'article

Référence. Sanger, W., de Marcellis-Warin, N. et Warin, T. 2019. Text-As-Data Analysis of Populist Parties Versus Government Parties : To Blend Or Not To Blend ?. Proceedings of the National Academy of Sciences of the United States of America. Soumis.

Le quatrième et dernier article de cette thèse change de terrain d'analyse, laissant de côté les données issues de Twitter. L'unité d'étude reste les élections, mais les données primaires concernent les manifestes de partis politiques européens. Depuis les deux dernières décennies, les différents pays d'Europe observent le succès électoral de partis politiques d'extrême droite, et de manière générale la montée du populisme.

En utilisant une méthodologie basée sur la science des données, l'objectif de l'article est de mesurer le rapprochement entre les doctrines de partis politiques et les thématiques associées à l'extrême droite en Europe. Plus de 600 manifestes de partis politiques sont analysés à partir des données en langue originale du Manifesto Project. L'analyse de similarité et l'identification de sujets au sein des textes permet de créer des données comparables pour l'ensemble des pays européens.

Une version préliminaire de l'article a fait l'objet de deux présentations orales effectuées en 2018 :

— Sanger, W. et Warin, T. (2018). Populisme et R : Analyse textuelle des manifestes politiques écrits en différentes langues. R à Montréal 2018. 5 juillet 2018, Montréal, Canada.

— Sanger, W. et Warin, T. (2018). Populisme et intégration européenne : une analyse en sciences de données des manifestes politiques, Colloque 623 - Sciences des données et sciences sociales : regards croisés, Responsables : Nathalie de Marcellis-Warin et Thierry Warin, 86ème Congrès de l'ACFAS. 8 mai 2018, Chicoutimi, Canada.

Une version retravaillée de l'article a été soumise à publication dans Proceedings of the National Academy of Sciences of the United States of America (PNAS).

Finalement, cet article de recherche a aussi permis de constituer une base de données concernant les indices de similarité entre les doctrines de partis politiques. Cette base de données a

fait l'objet d'un article spécifique, publié en avril 2019, dont les références sont les suivantes :

Sanger, W. et Warin, T. 2019. Dataset of Jaccard Similarity Indices from 1,597 European Political Manifestos across 27 Countries (1945-2017). Data in Brief.

7.2 Abstract

Populist parties have always existed. In this article, we look at whether populist parties have evolved recently to assimilate as government parties or on the contrary have distanced themselves from government parties. To capture the notion of populist parties, we limit ourselves to far-right parties as defined in the Manifesto Project. To capture the dynamics of the elements of language used by populist parties, we needed to limit our study to a field of interest. We chose to consider what the manifestos had to say about the European integration as a topic of choice. In order to measure the dynamics of the elements of language, we combine two very original methodological approaches based on Data Science techniques : first, we use a text-as-data approach based on the computation of Jaccard similarity indexes and second, we perform LDA analyses on political manifestos in their original language. We use a very unique database from the Manifesto Project starting in 2000 to match the birth of the Economic and Monetary Union.

7.3 Introduction

The goal of this article is to look at whether populist parties have evolved recently to assimilate as government parties or on the contrary have distanced themselves from government parties. To do so, we focus on the European integration since the year 2000. This article is at the junction between a research article and a method article. Indeed, it presents both original findings and new innovative study methods applied to new research questions.

The European integration process has encountered many challenges in the past decade, but it seems the ones ahead may be even more acute and important for its own future. Since the financial crisis - the biggest crisis for modern finance - Europe has had to face the sovereign-debt crises and was forced to implement a set of new regulations and institutions in order to provide a credible answer (the Banking Union) [Drakos et Kouretas, 2015]. The following sovereign-debt crises had re-opened the debates about the legitimacy and credibility of the European project, in particular the Stability and Growth Pact [Roman et Bilan, 2012]. Another challenge is an external shock having replications within Europe : the refugee crisis. The latter has been used as a way to promote the re-nationalization of political control. Furthermore and not the least of all challenges, is the Brexit vote on June 23, 2016, itself the result as well as the illustration of a strong populist movement in the United Kingdom (UK).

In academia, Europe has fostered a prolific research agenda across many academic fields. A query on Web of Science with the terms "european union" or "european integration" presents 6,512 articles from 1989 to 2018. By using VosViewer, a bibliometric visualization software [van Eck *et al.*, 2006], we mapped all keywords related to Europe. In figure 7.1, the landscape of the academic literature on Europe is provided.

A large portion of the academic research is related to trade and growth (lower left-hand corner), while a smaller proportion is about management and energy. We can find two parts : on the one hand on politics and integration and on the other hand, about the European institutions, Euroscepticism and identity. This terminology is the illustration in many ways of how Eurosceptic political parties reconsider the place of their respective countries within Europe. For instance, Brexit is a prime example, as well as the Lega Nord in Italy, the Front National in France or the Alternative für Deutschland in Germany.

In this article, we want to understand the dynamics of populist parties in Europe during the past 18 years (since 2000). To capture the notion of populist parties, we limit ourselves to far-right parties as defined in the Manifesto Project. To capture the dynamics of the elements of language used by populist parties, we needed to limit our study to a field of interest. We

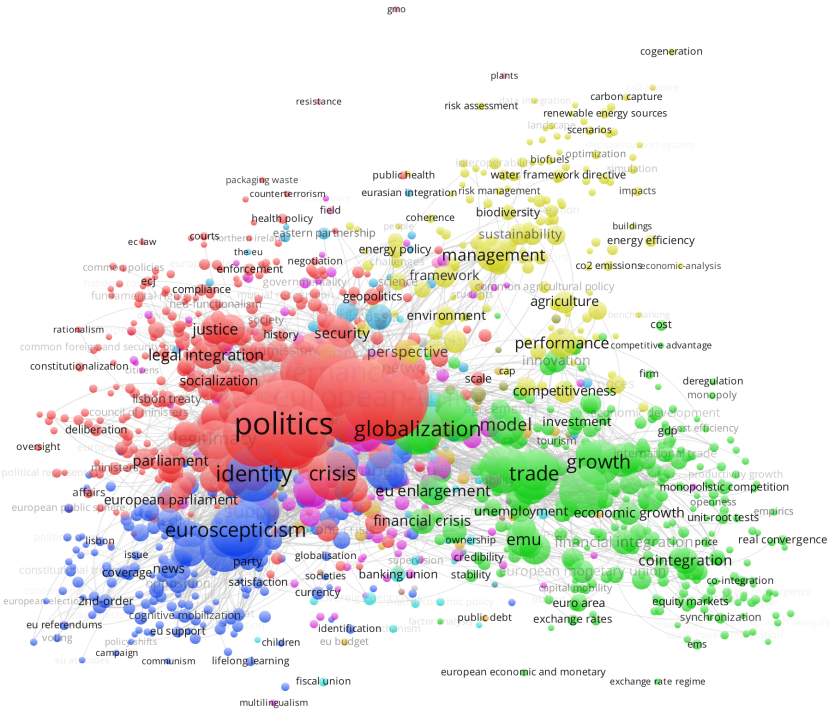

Figure 7.1 Landscape of the academic literature regarding Europe

chose to consider what the manifestos had to say about the European integration as a topic of choice. In order to measure the dynamics of the elements of language, we combine two very original methodological approaches based on Data Science techniques : first, we use a text-as-data approach based on the computation of Jaccard similarity indexes and second, we perform LDA analyses on political manifestos in their original language. We use a very unique database from the Manifesto Project starting in 2000 to match the birth of the Economic and Monetary Union. To do so we use a methodology based on data science to structure raw texts from political manifestos and create new variables. The rest of the article is organized as follow : the literature review is dedicated to political parties positioning as well as far-right parties and populism. After detailing our research question into two sub-questions, we explain our methodology based on text analysis. The results are presented in Section four and discussions are in Section five. The complete computations and visualizations are provided in the Appendix concerning each country (for Jaccard similarity and LDA) [1].

1. The complete Appendix is located at
https ://figshare.com/articles/toblendornottoblend_pdf/7781051/2

7.4 Literature Review

7.4.1 Political Parties Positioning, Branding and Lack of Information

In *An Economic Theory of Democracy*, [Downs, 1957] lays out a series of propositions in order to explain how political elements act in a society structured by the electoral process. From the voters' point of view, ideologies are being used to identify the elements related to political parties. These ideologies are the focal points in a context where the information costs to remember every detail related to a political party are too high. These information costs are well documented in the literature and have led to highlight cues used by voters.

[Kayser et Peress, 2012] studied how citizens base their decisions related to their country in comparison to foreign countries. Taking growth and unemployment as units of study, voters do hold politicians responsible, but only for the local dimension. [Bartels, 2005] focused on the 2001 Bush Presidency to suggest that causal links are not completely understood by individuals. Indeed, while most citizens do recognize social disparities, not all of them disapproved tax reductions that benefited wealthy people. This lack of information could systematically serve political parties. Indeed in the Canadian context, [Blais *et al.*, 2009] explain that the Liberal Party of Canada benefits from uninformed voters due to the party's notoriety. An improvement of the level of information can also change the voters' perception [Luskin *et al.*, 2002]. Credible signals can lead naive voters to vote for a specific party, compensating for a lack of information [Lupia, 1994].

In this context, the 24 th proposition from [Downs, 1957] stipulates that "political parties tend to maintain ideological positions that are consistent over time unless they suffer drastic defeats, in which case they change their ideologies to resemble that of the party which defeated them". This triggers a conflict between winning an election or focusing on core values throughout the party's history. When a shift in the electorate occurs, how do political party react ? Is it by positioning themselves differently, thus leading the way to the emergence of new political parties (for example in France with La République en Marche) or by emphasizing a fringe of an existing party (for example the Tea Party within the Republicains in the United States) ?

This is mainly the question that [Ezrow *et al.*, 2011] addressed in their article. Do political parties change position through time based on the reaction of their supporters or because the median voter has shifted ? They found that traditional parties are more prone to the change of the mean voter, whereas smaller political parties are more sensible to the change of their electoral base's opinion. [Downs, 1957]'s third proposition is that when two parties are against each other, they show convergence towards the median voter. [Ansolabehere

et al., 2001] attenuated this position by presenting an extensive study ranging from 1874 to 1996 focusing on the effect of polarization. The more an electoral district is polarized, the less the candidates converge. Parties are influenced by two driving forces : a centrifugal force that guides them towards an extreme position and a centripetal force that lowers their positions [Calvo et Hellwig, 2010]. The first force is more associated to smaller parties while traditional parties tend to be more stable, as observed through 13 different countries.

7.4.2 Populism and Far-Right Parties

In this context, recent elections have been characterized by strategies led by so-called populist parties. The literature about populism highlights two geographical areas (figure 7.2), Latin America and Europe. By observing the keywords associated to some 2 579 articles indexed on Web of Science (1989 - 2018), the elements related to Europe are about Euroscepticism, mass-media and Twitter, extreme right parties, far-right, islamophobia and immigration.

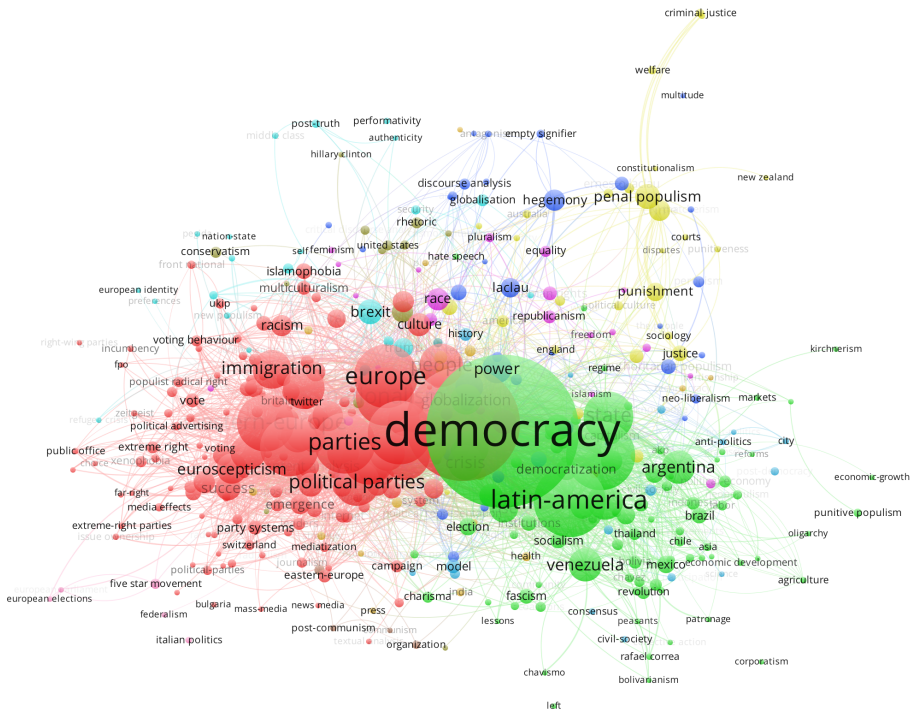

Figure 7.2 Landscape of the academic literature regarding populism

Scholars have categorized populism in many different ways and sometimes it is indeed challenging to precisely define populist parties due to the diversity of movements [Golder, 2016, Mudde, 2010]. These political parties are sometimes defined as protest parties [Steiner, 1997],

voters casting a protest vote during an election. However, a few common features should be highlighted. These parties promote radicalism [Mudde, 2010], through a profound change in the economic and political systems. Moreover, a Manichean vision of the world could be associated to their propositions [Eatwell, 2000] since those parties identify two groups of individuals. It is "them" vs. "us" [Mudde, 2004] where the ultimate goal is the interest of the people [Mazzoleni *et al.*, 2003]. "Them" could refer to the "elites" (traditional parties, intellectuals, media, wealthy individuals) or to groups based on their ethnicity, culture or religion [Betz et Johnson, 2004]. This last element sheds light to the concept of nationalism, which could be civic and inclusive or ethnic-centric and exclusive [Mudde, 2000]. However, such propositions are oriented towards the benefits of the "native" citizens of the country [Betz, 2004]. [Reynié, 2013] defines the populism in Europe as a "patrimonial populism", which could be characterized by a political offer that relies on individuals' anxiety, both cultural and material.

Recent electoral successes in Europe have emphasized the importance of populist and far-right political parties. In parallel, individuals in Europe are less and less participating in the electoral system in Europe, as can be seen in figure 7.3. While results vary through countries, the overall trend is a decline in the electoral participation in the European Parliament elections since the 80's. Turnout was a little higher than 60% at that time, but slowly diminished to reach around 40% in 2014. Countries recently admitted in the European Union (Czech Republic, Slovenia, Croatia, Hungary, Slovakia, Poland, Latvia, Estonia, Romania, Bulgaria) have turnout statistics lower than the European Union's average.

This is one of the manifestations of the loss of confidence in democratic institutions. Populist parties (particularly extreme right-wing parties) are seeking to challenge the institutions (European Union, governments, European Central Bank). In such a context, it is interesting to compare the success of populist parties with the lack of electoral participation ; would populist parties have such a success if the abstention would be lower ?

7.4.3 Research Question

Based on this literature, two strategies from far-right political parties could be formed. Do they emphasize their positions or try to resemble government parties ? The difficulty of characterizing such parties is also present in the literature, by the simple fact that when considering Europe, 28 countries are concerned, with 28 different political landscapes. Even though certain topics are more affiliated to populist parties, a comparison across European parties is lacking. Hence, our research question is stated as follows :

RQ : What are the dynamics of far-right political parties in Europe since 2000 ?

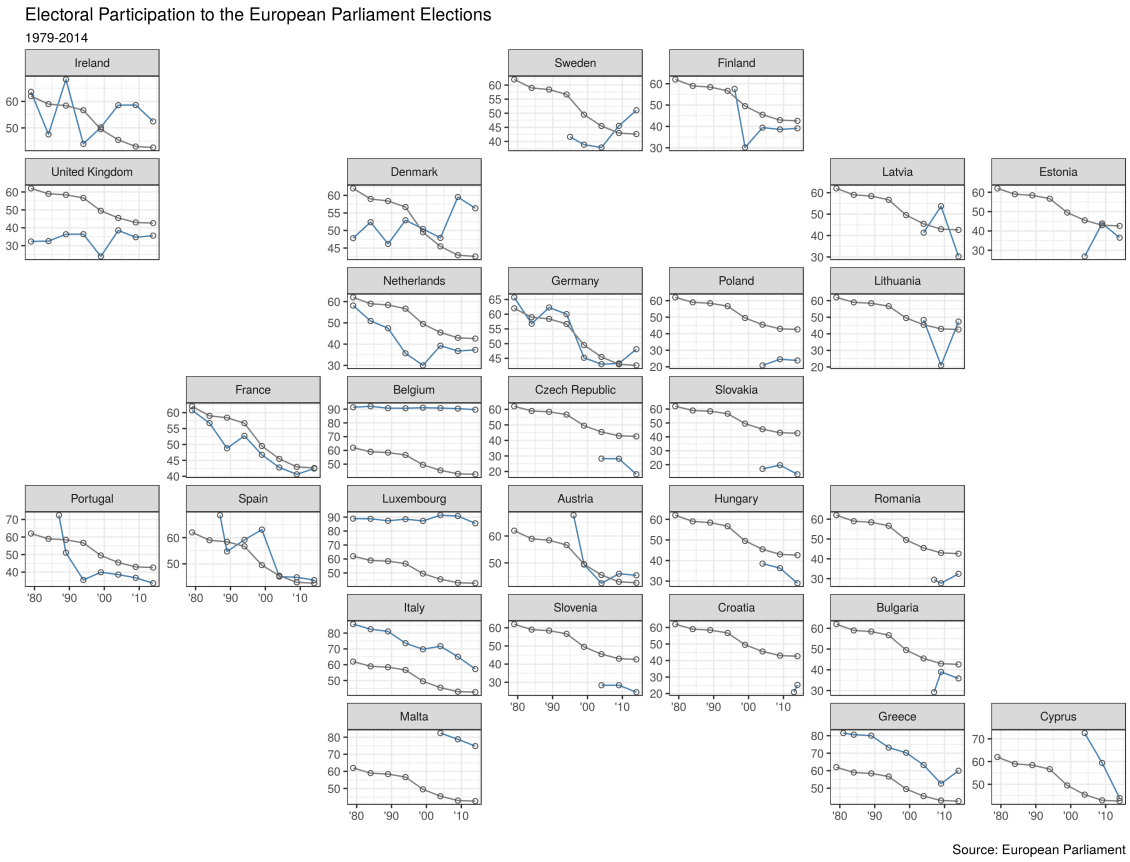

Figure 7.3 Electoral Participation to the European Parliament Elections

This broad research question can be divided into two parts.

— **RQ1** How close/divergent political platforms of far-right parties are from government parties ?

— **RQ2** What are the topics promoted by the electoral platforms of far-right political parties in Europe ?

Based on Data Science techniques, our goal is to answer those sub-questions by creating new indexes for text similarity between political platforms, as well as revealing topics that may be similar across countries. The next section will detail our approach as well as the data used for this research.

7.5 Methodology

7.5.1 Data

Having access to primary sources of data is of the utmost importance when developing a textual analysis-based methodology. In fact, raw text could be structured through Data Science techniques. These techniques have been used to extract policy positions [Grimmer et Stewart, 2013, Laver *et al.*, 2003]. We use the Manifesto Project's database to have access to the raw versions of texts from political parties [Lehmann *et al.*, 2018]. Overall, we consider 27 countries (Malta being excluded due to not being in the database). Indicators have been developed by several scholars since the inception of the Manifesto research project, which served as primary data for several hundred research articles. This database had been used to map policy preferences in 25 countries since 1945 [Budge et Klingemann, 2001] and since 1990 in Europe, the OECD and Eastern Europe [Klingemann *et al.*, 2006] for example. Moreover, the R library manifestoR [Lewandowski *et al.*, 2018] offers a convenient access, through an API, to the original text in the original language of the political parties.

This is precisely how we gather our database for this article. We first considered all manifestos available in Europe since 2000 until 2017, which reached a total of 676 manifestos. In total, 12,041,159 words were compiled to produce similarity indexes that will be detailed in the next subsection of the methodology section. However, for one part of the research, not all far-right political parties' manifestos are available in the Manifesto Project database. The selection of political parties is based on [Mudde, 2010] and [Golder, 2016].

For our article, the following political parties are considered. **Austria** : Freiheitliche Partei Österreichs (2003), Freiheitliche Partei Österreichs (2006), Freiheitliche Partei Österreichs (2008), Freiheitliche Partei Österreichs (2013) ; **Belgium** : Vlaams Belang (2004), Vlaams Belang (2007), Vlaams Belang (2010) ; **Bulgaria** : Attack (2009), Attack (2013), Patriotic Front (2014), Attack (2014), United Patriots (2017) ; **Croatia** : Hrvatska stranka prava (2000), Hrvatska stranka prava (2003) ; Denmark : Dansk Folkeparti (2001), Dansk Folkeparti (2005), Dansk Folkeparti (2007), Dansk Folkeparti (2011) ; **Estonia** : Eesti Konservatiivne Rahvaerakond (2015) ; **Finland** : Perussuomalaiset (2003), Perussuomalaiset (2007), Perussuomalaiset (2011) ; **France** : Front National (2002), Front National (2012), Front National (2017) ; **Germany** : Alternative für Deutschland (2013), Alternative für Deutschland (2017) ; **Greece** : Chrysi Avgi (2012), Chrysi Avgi (2015) ; Hungary : Jobbik Magyarországért Mozgalom (2014) ; **Italy** : Lega Norte (2001), Lega Nord (2006), Fratelli d'Italia (2013), Lega Norte (2013) ; **Latvia** : Tevzemei un Brivibai/LNNK (2006), Nacionala apvieniba (2010), Nacionala apvieniba (2011), Nacionala apvieniba (2014) ; **Netherlands** : Partij voor de Vrijheid

(2006), Partij voor de Vrijheid (2010), Partij voor de Vrijheid (2012) ; **Slovakia** : Slovenská národná strana (2006), Slovenska narodna strana (2010), Slovenska narodna strana (2012) ; **Slovenia** : Slovenska Nacionalna Stranka (2004), Slovenska Nacionalna Stranka (2008), Slovenska Nacionalna Stranka (2011) ; **Sweden** : Sverigedemokraterna (2010), Sverigedemokraterna (2014) ; **United Kingdom** : UK Independence Party (2001), UK Independence Party (2015).

7.5.2 Text Similarity

Obtaining a metric of text similarity is possible through a set of measures. Amongst them is the Jaccard similarity coefficient, which compares the set of common items of two entities to the set of unique items of these two entities, such as :

$$J(set_1, set_2) = \frac{|set_1 \cup set_2|}{|set_1 \cap set_2|} \tag{7.1}$$

with set_1 and set_2 two text to be compared. However, the Jaccard similarity ratio was used in the academic literature more than a century ago in another research field : botany. In the early 1900s, Paul Jaccard, then Professor in Lausanne (Switzerland), was analyzing how flower species were distributed between different areas in the Alps and in the Jura. He wanted to statistically measure the diversity of floral distribution in order to reveal and understand the patterns of ecological variations [Jaccard, 1902b]. A first set of measures between three territories were published in 1900 [Jaccard, 1900].

He found that nearby territories only share a relative low number of common species [Jaccard, 1901a]. One the other hand, for areas that are characterized by an apparent uniformity in their biodiversity such as Jura, Pr. Jaccard found that only 40% of herbal species are common between six areas jointly compared one to the other [Jaccard, 1901b]. Such measure of community was also used to compare an area of Jura and another one in the Alps [Jaccard, 1902a] : he found that these distanced areas shared more common species than two areas belonging to the same mountain side. Hence, using a ratio of similarity enables to compare different elements (here botanical species) from two foreign sets (areas between of two mountains) and at the same time from two similar sets (two areas within the same region).

By adopting a purely statistical approach, Pr. Jaccard found that similarity in ecological conditions of two nearby territories could be approximated by the similarity in their floral distributions ; the more diverse in ecological conditions two territories are, the less a certain number of floral species will be commonly shared [Jaccard, 1902b].

In international business, an analog approach has been used to study Free Trade Agreements (FTAs) notified to the World Trade Organization. In fact, [Alschner *et al.*, 2017] compared FTAs to the Trans-Pacific Agreement. From these values of similarity, they augmented the equations of gravity models to test the effect on trade of similar FTAs. The same approach was used to map the 2,100 different international investment agreements through Jaccard Similarity based on text divided into n-gram 5 characters [Alschner et Skougarevskiy, 2016].

While being a language-dependent technique, the Jaccard similarity could be applied to different languages. [Alschner *et al.*, 2017] warned about comparing FTAs from the same language (English and Spanish FTAs). Besides the previously mentioned articles, the Jaccard coefficients have been used in other contexts for text classification of Arabic language [Al-Kabi et Al-Sinjilawi, 2007, Thabtah, 2008] and Thai language [Niwattanakul *et al.*, 2013]. Moreover, [Lodhi *et al.*, 2002] stressed that bag-of-words techniques do not reflect the order of words into sentences. By using n-gram 5 characters, such limitation is avoided.

It is worth mentioning that other techniques of text similarity are being used to compare texts between each other, especially based on term frequencies, such as TF-IDF (term frequency-inverse document frequency). Recently, using Twitter data, several methods have been used in order to try to identify the authors of the anonymous op-ed published in the New York Times of September, 5th of 2018. Three approaches will be described. TF-IDF serves as input data for computing cosine similarity between the op-ed tf-idf and several senior members' Twitter time lines of the Trump administration [Robinson, 2018]. A second methodology involves considering 107 numeric features for each segment of the op-ed [Kearney, 2018a]. Those features were then averaged and compared to Twitter time lines of senior members of the Trump administration through a measure of correlation. A third proposition was made by [Misra, 2018] who used tokenized versions of tweets to compute Burrows' Delta in order to attribute the authorship of the op-ed. The three methodologies rely on [Silge et Robinson, 2017]'s approach to treat text as data and transform it as tidy text before any analysis. Those computations were made through the R programming language.

Here, the methodology of comparing two texts is also made through the R programming language. First, as suggested by [Alschner et Skougarevskiy, 2016], each individual text is segmented into n-gram 5 characters. The segmentation is done five times, each repetition with a starting point shift by a character. We use the function jaccardsimilarity() from the R library textreuse [Mullen, 2016] to evaluate the percentage of similarity between two different texts. The complete methodology is explained in [Sanger et Warin, 2018b], and the complete database of Jaccard similarity indexes is provided in addition to this article.

7.5.3 Topic Modelling

Topic modelling is a methodology widely used in Comparative Politics. Human codification of sentences could be done, but drawbacks such as time consumption, redundancy of tasks and limitations regarding the methodology are still present [Quinn *et al.*, 2009]. The use of computer-centric methods tries to reduce the effects of these drawbacks. For example, [Quinn *et al.*, 2009] developed a methodology that only takes into account the number of topics within a text for the automatic coding of the Congressional record for the 105 th to the 108 th U.S. Senate. When comparing five methods of discrete text categorization, topic modelling appears as the methodology with the fewer pre-analysis and analysis costs (compared to reading, human coding, dictionaries and supervised learning). Because automatic classification methodologies can achieve a level of classification similar to human coders with no drawbacks. Indeed, thanks to access to powerful and cheap computer programs, programs are set to outdo human coders in the long run [King et Lowe, 2003].

In conjunction with the Manifesto Project's database, topic modelling have been used in Comparative Politics. The database offers an access to a rich amount of complex data, which can be leveraged to compare political parties' positions across different countries. [Jankowski et Gross, 2017] have produced a complete text analysis of local political manifestos in Germany using Wordscores and Wordfish methods to estimate party positioning. [Zirn *et al.*, 2016] developed and tested the accuracy of a methodology aimed to automatically classify each sentence from the Manifesto Project to specific topics of the database. By using manifestos of three U.S. Senate elections (2004, 2008, 2012 from both the Republicans and the Democrats), they achieve an automatic classification of almost 80% of the topics. [Glavaš *et al.*, 2017] dealt with multiple languages from the Manifesto Project database by transferring all text to a reference language, i.e. English.

Our methodology regarding topic modelling relies on several steps, including preprocessing data extracted from the Manifesto Project's API. First, raw text are tokenized, i.e. each word is uncapitalized, numbers are stripped from original texts and punctuation is removed. Using the function unnesttokens() from the tidytext library [Silge et Robinson, 2017], the dataframe containing the processed data has now a single word per row. With the help of a regular expression (regex) [: *alpha* :]2,+, each word that is less than two single non digit characters is not considered. In addition to this, stopwords from each languages are removed, assessed by the function stopwords() of the tm library [Feinerer *et al.*, 2018] and the library stopwords [Benoit *et al.*, 2017].

The case of France will be taken as an exploratory example, but the methodology will be carried on to the other countries as well. We will focus on two specific elections, in 2002 and

in 2017. For those two years, the Front National has been present in the second round of the French Presidential election. We want to explore the most preferentially used words within this party's manifestos while comparing with the parties that won the election (UMP for the 2002 election and En Marche for the 2017 election).

We will compute the tf-idf for each manifesto. The tf-idf of a word i from a document j is calculated as follows :

$$tfidf_{i,j} = tf_{i,j}.idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}}.log\frac{|D|}{|d_j : t_i \in d_j|} \tag{7.2}$$

with $n_{i,j}$ the number of occurrences of a word i from the document j divided by the total number of words from this document ($\sum_k n_{k,j}$). This frequency is weighted by the logarithm of the number of documents $|d_j : t_i \in d_j|$ of a corpus D where the word i appears.

Regarding the divergence/convergence of electoral platforms, in completion of the Jaccard similarity coefficient, a Pearson's product-moment correlation will be assessed taking into account the distribution of words.

Finally, we perform a Latent Dirichlet Allocation (LDA) to model the topics associated to far-right political parties in Europe. The methodology is based on [Silge et Robinson, 2017] who used a tidytext approach to model topics in texts. The goal of this methodology is to clusterize words from political manifestos into the most prevalent topics and to compare them in Europe since 2000. First, for each political manifesto, we consider words that are found more than 5 times. We use the variational expected-maximisation algorithm [Blei *et al.*, 2003] from the R library topicmodels [Hornik et Grün, 2011] to segment our database into a specific number of topics. The number of topics within a text is a parameter set by the researcher. In [DiMaggio *et al.*, 2013] it was set to 78 and 82 topics ; in [Fligstein *et al.*, 2014] to 23 and 24 ; in [Quinn *et al.*, 2009] it was 42. In our case, we set the number of topics at 20 due to the high number of categories in the Manifesto Database. Indeed, if we would use all categories, since some of the manifestos are quite short, it would be harder to properly differentiate concepts and topics. However, by using [Asuncion *et al.*, 2009]'s perplexity measure, it is possible to optimize the number of topics within a text.

7.6 Results

7.6.1 Similarity between Political Manifestos

For each country, political manifestos were compared pair-wise. Figure 7.4 showcases the results obtained for France for the past two elections (2012 and 2017). A few elements can

be noticed in this figure. First, UDI and Les Républicains have the highest value of Jaccard similarity in 2017. This is due to the fact that both political parties shared the same electoral platform and worked together during the election. Another high index of similarity is between the MoDem's manifesto in 2012 and the MoDem's manifesto in 2017. The similarity reaches 81%, meaning that most of the elements from one election were transmitted into the platform of the next election. The Front National's manifesto of 2017 shares the highest resemblance with the UDI/Les Républicains' of 2017 (34%). A similar index is found between two subsequent electoral platforms (2012 and 2017) of the Front National. On the opposite spectrum, the two political formations that share the least with the Front National is the Parti Socialiste (20%) and the Parti Communiste (10%).

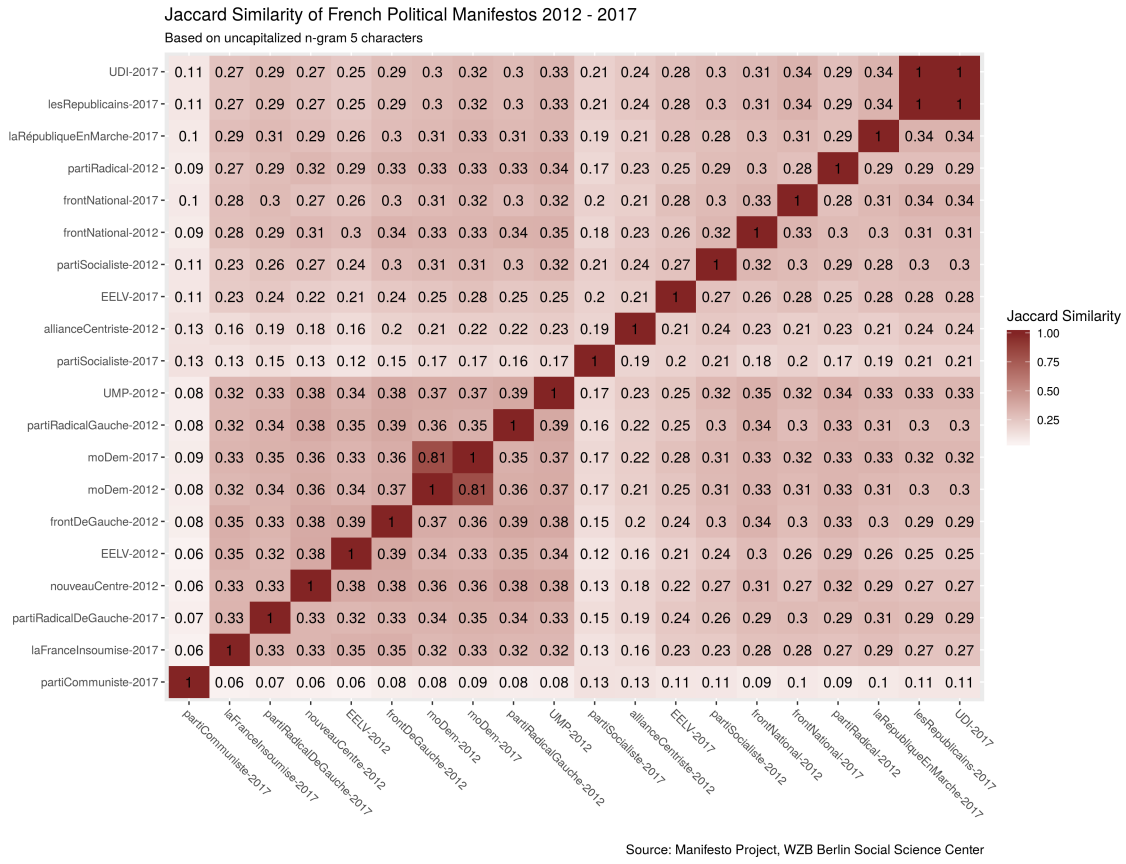

Figure 7.4 Heatmap of the Jaccard similarity indexed of France (2012-2017)

In the Appendix, the complete heatmaps are presented by country in order to obtain a complete coverage of the similarities between European political parties since 2000.

7.6.2 Similarity between Far-Right Parties and Government Parties

For each election provided in the Manifesto Project's database, we computed the similarity index between government parties and far-right parties. Not all elections have been covered since not all political manifestos are available in the database. However, 48 elections across 18 countries are analyzed. Figure 7.5, provides an overview of the similarity evolution in Europe.

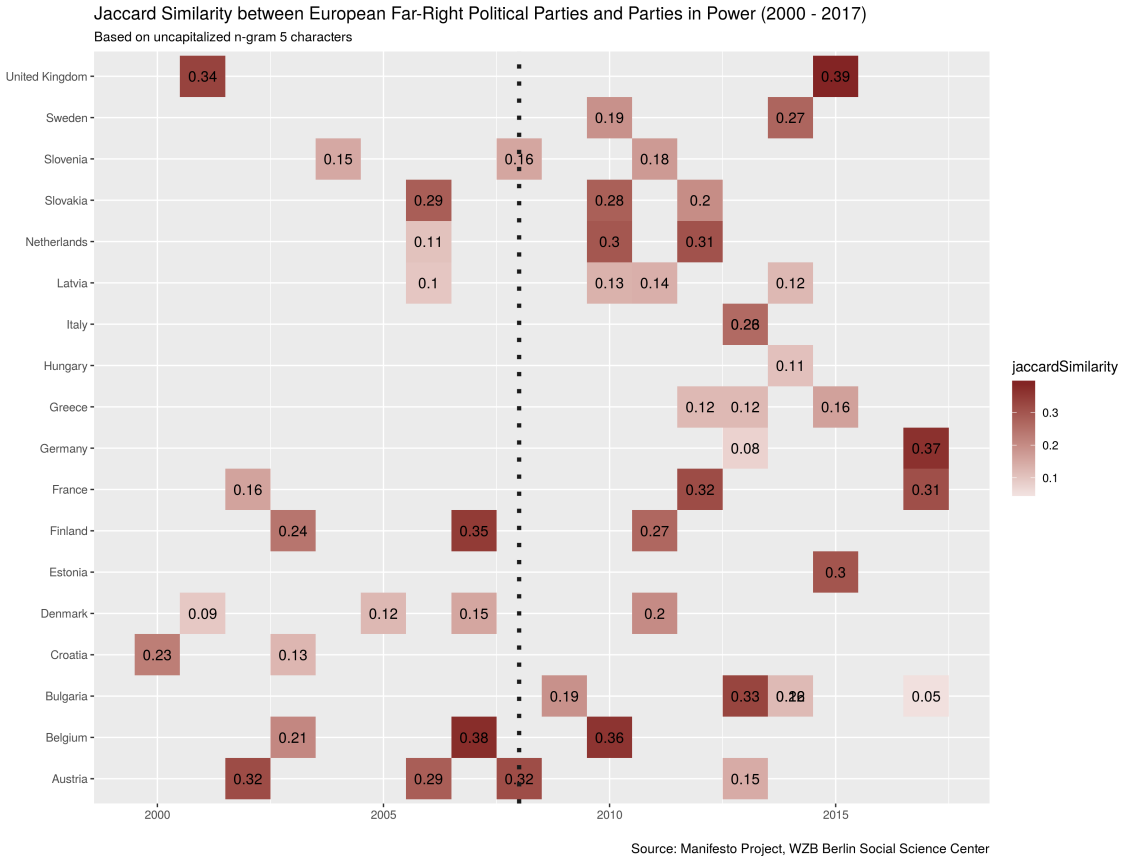

Figure 7.5 Evolution of similarity through Europe since 2000

Three groups of countries could be identified. First, the countries where far-right political platforms share more than 30% of the content of government political parties (Austria prior 2010, Belgium, Estonia, Finland, France, Germany, Netherlands and the United Kingdom). A second group could be highlighted with similarities from 20% to 30% (Italy, Slovakia and Sweden). The rest are countries where government political parties share the least common elements with far-right parties (less than 20%) : Bulgaria, Denmark, Greece, Hungary, Latvia and Slovenia.

7.6.3 Topics of European Far-Right Parties

In this section, before considering all the countries in our database, we focus on a specific case-study : two elections in France will be studied. In 2002, the Front National has been present during the second round of the Presidential election, which was received with great surprise since it was the first time this political party had such an electoral success. Jean-Marie Le Pen party's was against the incumbent candidate and former President, Jacques Chirac from the Union des Mouvements Populaires. 15 years later, the year 2017 saw a similar result, with the Front National, now led by Jean-Marie Le Pen's daughter Marine Le Pen, sparring the newly formed party of Emmanuel Macron, La République en Marche.

The question remained to explain how the topics of these different political formations have evolved across 15 years. To do so, we present, in the next two figures (7.6 and 7.7), the distribution of words between on the one hand the manifesto of the Front National of 2002 and three other parties (Union des Mouvements Populaires 2002, La République en Marche 2017 and the Front National of 2017). One the other hand, the manifesto of the Front National of 2017 is compared to the ones of the Front National of 2002, the Union des Mouvements Populaires of 2002 and La République en Marche of 2017.

Words are scattered around a dashed line in these figures. A word located on the dashed line means that the word is equally shared between the two political manifestos in terms of frequency. If a word is at the left of the dashed line, it is more present in terms of frequency in the referred political manifesto (FN 2002 or FN 2017) than in the compared ones. Several thousands words are considered in such visualization. Let's focus on two specific ones, "Europe" and "France."

In 2002, "Europe" has taken 0.0429% of the length of the political manifesto of the Front National, compared to 0.0107% for the Union des Mouvements Populaires. In 2017, La République en Marche focused his message towards the place of France within Europe, which could be observed in the proportion of the word Europe with a value of 0.0536%. The Front National of 2017 dedicated only 0.00357% of his manifesto to this topic. In order words, the far-right political party of 2002 focused more on Europe than in 2017 strictly considering the frequency of this word, compared to the electoral winning of the legislative election.

When considering the word "France", the Front National of 2002 use 0.49% of the length of its manifesto towards the name of the country, compared to 0.0214% for the Union des Mouvements Populaires. This difference has been lowered 15 years later, with 0.129% for the Front National of 2017 and 0.0751% for La République en Marche of 2017.

What are the words used preferentially by each political party ? It is expected that political

Figure 7.6 Term frequency comparison between UMP 2002, FN 2017, LREM 2017 and FN 2002

platforms share a certain degree of similarity since they are written towards the same election. However, political parties may want to emphasize topics that they will consider to be their focus.

The term-frequency inverse-document-frequency (TF-IDF) provides an overview of the words most used by a political party relative to the terms used by all political parties. Here, we compute in figure 7.8 the words that appear relatively more often in the political platforms of the Front National (2002, 2017) as well as in the government parties of 2002 (UMP) and 2017 (En Marche).

Amongst the words used preferentially by the Front National of 2002, **nationals** ("ressortissants"), **to prefer** ("préférence") and **nuclear** ("nucléaires") emerged as the most important. His opponent that year, the UMP, emphasized words related to **the State** ("État") as well as several action verbs such as **to adopt** ("adopter"), **to help** ("accompagner"), **to assure** ("assurer") and **to consolidate** ("consolider").

Figure 7.7 Term frequency comparison between UMP 2002, FN 2002, LREM 2017 and FN 2017

Fifteen years later, the focus of the Front National seemed to have shifted. In fact, words like **to sustain** ("maintenir"), **startups**, **clubs**, related to **the bank industry** ("bancaires") or **heritage** ("patrimoine") were more present. On the contrary, the tone of En Marche's manifesto is different compared to the three other ones. It is the only party where verbs are conjugated at the first person of the plural in the future tense (those verbs appear preferentially in En Marche's manifesto compared to the other ones'). We will **create** ("créerons"), we will **do** ("ferons"), we will **put** ("mettrons"), we will **give** ("donnerons"), we will **reduce** ("réduirons"), we will **suggest** ("proposerons"), we will **develop** ("développerons"), we will **build** ("construirons") are most preferentially found in this year's manifesto.

In order to complete the comparison between political platforms, a Pearson's product-moment correlation is done between the correlation of word frequencies of political platforms. The correlation result between the Front National of 2002 and the Union des Mouvements Populaires is about 0.786. This a slighly lower in 2017, with 0.732 between the Front National and La République en Marche. Interestingly, the correlation is the highest one between the

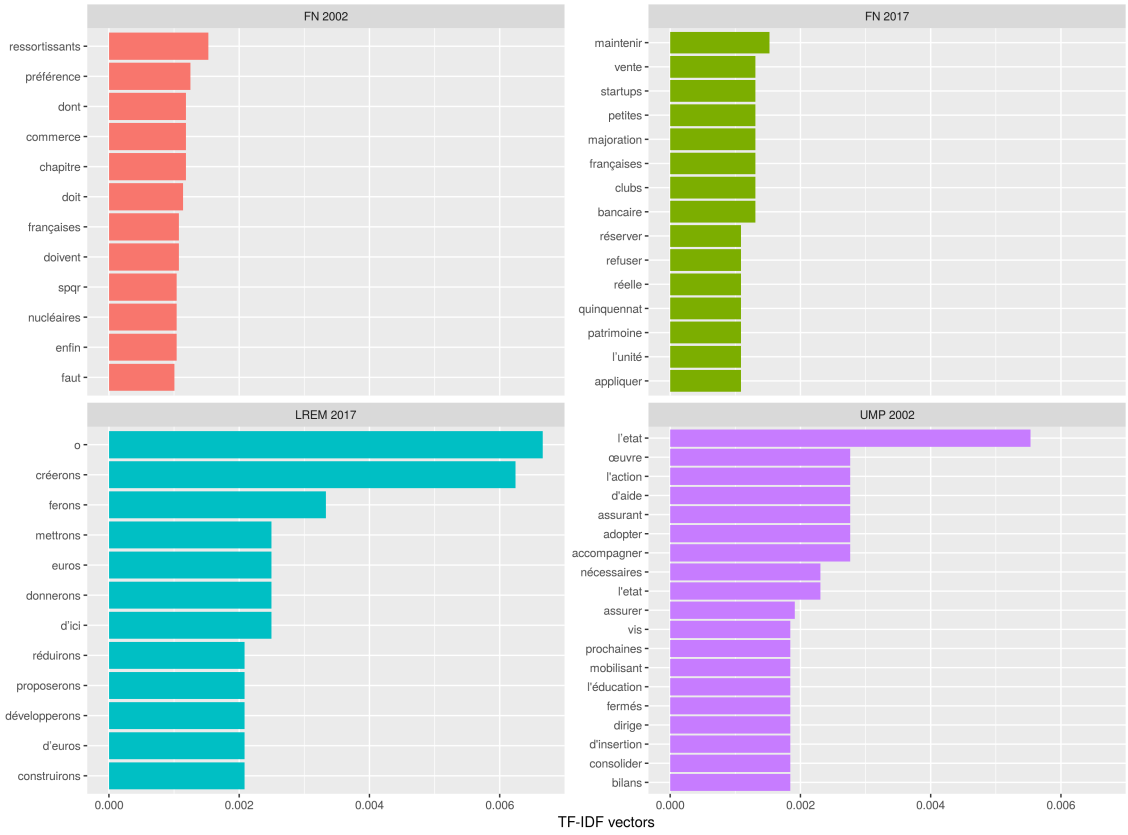

Figure 7.8 Tf-idf between LREM 2017, FN 2002, UMP 2002 and FN 2017

two political platforms 15 years apart, i.e. between the Front National of 2002 and the Front National of 2017.

The complete comparison of political platforms between far-right parties and government parties is assessed through Pearson's product-moment correlations, located in the Appendix.

Finally, we perform an LDA analysis on the 20 topics promoted by the Front National since 2002. We take into account three political manifestos (2002, 2012 and 2017) that are available in the database. For each of the 20 topics, we present in figure 7.9 the 10 top terms to have a glimpse of these categories.

The first topic is related to the notion of **work** (travail), **territory** (territoire) and **people** (personnes). The concepts of **immigration and culture** (topic 7) or **immigration and working compensation** (topic 12) appear to be important themes. **Europe** is across topic 4, **security** topic 13. However **companies** (entreprises) is located on several topics (2, 6, 7, 9, 11, 13, 14 and 19). For other countries, LDA analyses are performed and presented in the Appendix.

7.7 Discussion and Conclusion

Do smaller parties tends to look like already established political structures or do they tend to stay original ? This question was the main goal this article. By offering indexes helping structuring raw text in different languages, this article tried to answer the first sub-question, which was aimed to characterize the divergence or divergence of far-right political parties. Not surprisingly, the answer varies across countries. In the case of France, it is interesting to note a convergence between the Front National and the government parties.

Smaller or bigger similarity results between far-right political parties and government parties do not guarantee that those parties differentiate from each other in a specific way. To be more precise, a potential shift could be explained by a change in the content of text from both parties, i.e. traditional parties might radicalize their message also. The direction of such movement is not yet well captured and show be the focus of a next study. The several steps of our methodology offer a refined look at the textual content of political manifestos in a comparative approach. Even though the texts considered are published in different languages, those texts can be compared through common indexes, hence characterizing how European political parties are branding themselves across Europe.

The methodologies developed in this article could be used in two specific ways : (1) as showcased with the example of the Front National, it is possible to follow the ideology of a party and how it evolves through time. What is the impact of a change of leader over the party line and communications ? (2) This methodology could be adapted to specific topics of interest. Instead of using entire manifestos, we could isolate paragraphs or texts dedicated to a theme (for example the environment) and study what are the main lexical positions of several political parties during an election or through time.

One limitation of our research is the fact that we rely on data from the Manifesto Project's database. While being a source of rich and precise information, not all political manifestos are stored in the database, even though the coverage of the project tends to increase over the years. A second limitation to our study comes from the LDA methodology. Two elements could be improved in future works. First, the number of topics should be adapted specifically to the corpus of texts. This parameter, k, can be optimized through computations that are resource consuming but might highlight dedicated groups of elements. As stated in the methodology section, [Asuncion *et al.*, 2009] offers an improvement of the LDA methodology by identifying the appropriate number of topics within a corpus of texts. The second improvement of the LDA methodology is to consider the LDA through time. How did the topics promoted by the different political parties tend to evolve since 2000 ?

For future works, the indexes that were developed in this study should be used in conjunction with mainly three sets of data. On the one hand, the electoral success of political parties could be added to the database through either the number of electoral seats won or by the amount of votes in percentage. On the other hand, the Manifesto Project's database gathers an impressive amount of indicators, some of them related to the European Union and European Integration in positive (per108) or negative (per110) terms. In addition to this, the Eurobarometer and the future European elections of 2018 would serve as primary sources of data to characterize populism or the view of the European Union across the countries.

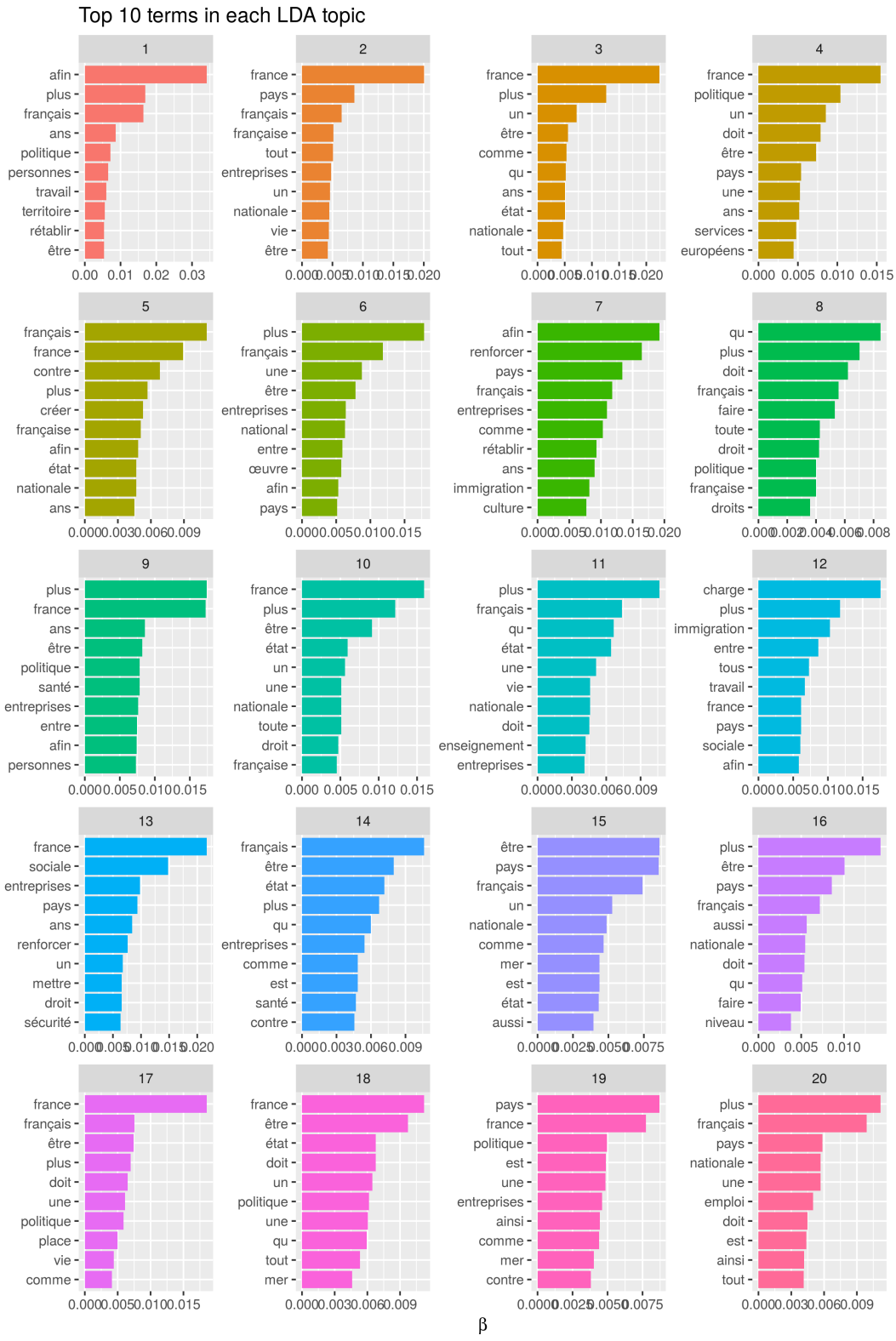

Figure 7.9 LDA analysis for the Front National (2002, 2012, 2017) with 20 topics

CHAPITRE 8 DISCUSSION GÉNÉRALE

Pour synthétiser le corps de la thèse, cette section abordera une discussion générale organisée en trois parties. Tout d'abord, un résumé des quatre articles de recherche ayant structuré cette thèse sera proposé. Par la suite, nous présenterons les apprentissages tirés du travail doctoral. Finalement, les perspectives de recherche futures seront étayées.

8.1 Synthèse

La revue de littérature a permis de délimiter un champ de recherche en effervescence d'où émergent de très nombreuses publications. Que ce soit concernant la politique et les élections, les données massives, ou l'intersection de ces deux disciplines de recherche, la forte progression du nombre d'articles publiés dès le début de la première décennie des années 2000 montre l'intérêt d'un champ dont les thématiques de recherche et les méthodologies de recherche sont encore en évolution.

Nous avons cherché à comprendre les liens entre citoyens, partis politiques et institutions dans le cadre démocratique des élections. Les lacunes d'interprétation de l'information pour un électeur est un élément majeur, puisque ces lacunes se verraient amplifiées à une époque de boulimie informationnelle. L'utilisation des médias de masse (à travers les journaux, la radio, la télévision et depuis le milieu des années 2000 Internet et les réseaux sociaux) montre que les liens de cause à effet pour les électeurs ne sont pas évidents, surtout en contexte électoral. Les nouvelles sources de données que sont les données issues des médias sociaux offrent des défis méthodologiques majeurs, puisque les questions de représentativité des échantillons sont au centre des études liées aux prédictions électorales. Les partis politiques, comme les citoyens, produisent des données liées à leurs comportements, mais aussi peuvent se servir de ces données. Dépassant le cadre des sondages, ces données massives deviennent donc une fenêtre sur l'étude de comportements politiques.

Au début de cette thèse, une question de recherche principale avait été proposée, mêlant science des données et données massives pour comprendre les processus démocratiques au XXIème siècle. Comment les données massives et la science des données peuvent être utilisées pour comprendre les processus démocratiques à l'ère d'Internet ? De cette question de recherche générale, quatre sous questions ont été étayées et répondues respectivement par chaque article composant le corps de la thèse.

1. Comment sont perçus les partis politiques au Québec sur les médias sociaux ?

2. Comment se transposent les dynamiques électorales sur les médias sociaux en contexte canadien ?
3. Comment pallier aux lacunes institutionnelles et au manque d'informations pour le suivi d'une élection ?
4. Comment utiliser la science des données pour caractériser la montée du populisme ?

La méthodologie repose essentiellement sur des méthodologies quantitatives en science des données, soit des analyses économétriques et des analyses textuelles basées sur des données massives collectées en temps réel. Les données utilisées reflètent la diversité et la quantité de ces données massives, avec des millions de tweets analysés au Québec, au Canada et au Nigeria, ou plusieurs centaines de manifestes politiques européens depuis le début des années 2000.

L'article 1 (chapitre 4) vise à comprendre comment les perceptions des différents partis politiques évoluent à travers les campagnes électorales actuelles. Deux hypothèses ont été utilisées : tout d'abord, il est possible de mesurer en temps réel la résonance de sujets politiques sur Twitter, et ensuite que ces thématiques de campagne peuvent être associées préférentiellement à des partis politiques lors d'une élection. Pour cela, nous avons utilisé comme unité d'analyse les messages issus de Twitter publiés dans le cadre des élections provinciales du Québec en 2014.

Au cours de ces élections, le Parti Québécois dirigé par Pauline Marois déclencha une élection afin de tenter d'obtenir une majorité de sièges au Parlement québécois. Son pari échoua, et ce malgré une position de candidate sortante et des appuis politiques de taille dès le début de la campagne. Plus de 670 000 messages ont été récoltés pour comprendre comme les thématiques de campagne sont associées aux différents partis politiques. À travers des modèles logit, cette associativité de thématiques électorales aux partis politiques a pu être mesurée tout au long de la campagne électorale. Il est intéressant de noter que certains phénomènes liés à l'étude des médias sociaux ont été relevés, notamment le fait que la composante des données en temps réel offre une opportunité de mesure plus précise que des sondages traditionnels. Ce dernier élément sera repris à travers l'article 3 (chapitre 6) de la thèse concernant les élections au Nigeria.

Le deuxième article de la thèse (chapitre 5) prend appui sur l'article premier de la thèse et se concentre sur le contexte canadien. Ainsi, lors des élections fédérales canadiennes de 2015, comment se transposent les dynamiques électorales sur les médias sociaux ? À nouveau, deux hypothèses de travail ont été esquissées dans le cadre de cette étude. Dans un premier temps, les dynamiques électorales canadiennes se reflètent dans les messages publiés sur Twitter, et ensuite il est possible de mesurer la persistance de certains enjeux de campagne.

À travers cette élection, le Canada connu sa plus longue campagne électorale, s'étendant sur 80 jours au total. À l'image de la campagne québécoise, le candidat sortant, Stephen Harper du Parti Conservateur du Canada, ne réussit pas à conserver sa place à la tête du gouvernement. Malgré un début de campagne en troisième position dans les intentions de votes, ce fut le Parti Libéral du Canada, porté par Justin Trudeau, qui remporta les élections et put former un gouvernement majoritaire. En se basant sur l'analyse de près de 3,5 millions de tweets, deux méthodologies d'analyse textuelles ont été employées. Tout d'abord, la polarité des messages concernant chacun des chef·fe·s de partis politiques fut étudiée. Le Premier ministre sortant fut associé principalement à des messages à teneur négative, tandis que le Parti Libéral de Justin Trudeau connut des périodes plutôt positives. Surtout, une analyse des thématiques à travers une allocation de Dirichlet discrète (LDA) permit de mettre en évidence les différents sujets de la campagne électorale, notamment les scandales ayant eu lieu au lancement de cette période électorale.

Le troisième article de la thèse (chapitre 6) est le dernier utilisant comme données massives les données issues des médias sociaux. La question de recherche fut de savoir si les médias sociaux pouvaient pallier à des lacunes institutionnelles ou à un manque d'informations dans le cadre du suivi des élections. Ce faisant, nous nous sommes concentrés sur les élections ayant eu lieu au Nigeria en 2015. La première hypothèse de recherche fut de savoir si les médias sociaux pouvaient offrir des mesures d'analyses plus fines que des sondages traditionnels. La seconde hypothèse de recherche concerne un aspect méthodologique du travail de recherche, notamment la détermination de modèles économétriques adaptés au traitement des données provenant des médias sociaux.

Dans le cadre de cette élection, le manque d'informations officielles fut important. En effet, les instituts de sondages ne proposèrent pas des comptes rendus empreints d'impartialité, puisque les résultats varièrent au jour le jour dans des intervalles de confiance élevés. Ainsi, l'étude de deux bases de données de tweets dédiées aux élections nigérianes fut requise. Dans un premier temps, plus de 555 000 messages géolocalisés furent collectés puis placés sur le territoire nigérian. Cela permit de mettre en évidence des zones géographiques ayant une affinité pour un certain candidat, ou une analyse à travers certaines villes du pays. La seconde base de données de plus de 1,5 millions de tweets permit de mesurer l'associativité de thématiques électorales par rapport aux deux candidats en présence, soit le Président sortant Goodluck Jonathan et son challenger Muhammadu Buhari. Les modèles économétriques (logit binaire, logit multinomial, logit stéréotypé) permirent de prendre en compte les erreurs économétriques reliées à la nature des données issues des médias sociaux. Un élément à souligner lors de cette étude est le fait que des changements d'évolution de certains indicateurs (notamment la proportion de messages géolocalisés concernant un candidat) a pu être notée

quelques jours avant la fin de la campagne électorale. Ce phénomène est passé inaperçu à travers les sondages disponibles au cours de la campagne électorale.

Finalement, le quatrième article de la thèse (chapitre 7) se concentre sur un autre type de données massives, soit les données provenant des manifestes politiques des partis politiques en Europe. De manière générale, cet article répond à la question de recherche visant aux méthodes de mesure de la montée du populisme. À nouveau, deux hypothèses de recherche furent testées. Une première proposition fut d'observer l'ajustement de programmes électoraux entre les partis politiques à l'échelle européenne. La seconde proposition est de développer un cadre pour comparer les propositions et thématiques des partis d'extrême droite en Europe.

Un réel défi méthodologique caractérisa ce travail de recherche. En utilisant la base de données du Manifesto Project, un accès privilégié à des textes de partis politiques en langue originale a été possible. Toutefois, ces textes correspondent à des quantités importantes de données, avec pour l'entièreté de l'article plus de 12 millions de mots analysés, à travers 676 manifestes politiques de 27 pays différents entre 2000 et 2018. Deux mesures de similarité ont été utilisées, soit d'abord les indices de Jaccard et l'analyse des fréquences de distribution des différents mots. La première méthode prend en compte le contexte des mots au sein d'un texte tandis que la seconde ne se concentre que sur la fréquence des différents termes. Reprenant la méthodologie d'analyse textuelle du deuxième article (chapitre 5), des analyses de thématiques associées aux différents partis politiques d'extrême droite ont été effectuées. Ces analyses sont portées par des dictionnaires de mots adaptés aux différents langues utilisées à travers ces manifestes politiques. L'apport majeur de cet article est la création de nouveaux indicateurs permettant la comparaison de textes écrits à des époques différentes, et dans des lieux différents. Ainsi, à travers la science des données de manière générale, des mesures de correspondance entre partis politiques ont pu être établies, caractérisant ainsi le climat politique européen des vingt dernières années.

Les quatre articles de la thèse se répondent de plusieurs manières. De par la diversité des données utilisées et des techniques mises en oeuvre, ces quatre articles permettent de répondre à la question de recherche principale, soit de comprendre l'utilisation de la science des données et des données massives pour décrypter les processus démocratiques actuels. Bien que chaque article vise à répondre à une sous question de cette thématique de recherche, les méthodologies dépassent le cadre des articles, puisqu'utilisées de manière conjointe au fil de la thèse. C'est cette utilisation complexe de méthodes de structuration des données qui permit d'appréhender des dynamiques électorales et démocratiques moins bien définies comme le populisme.

8.2 Apprentissages concernant une campagne électorale

Tandis que la partie précédente fait état des réponses apportées par chacun des articles à la question de recherche principale, plusieurs ponts peuvent être tissés entre les différentes parties de la thèse. De plus, plusieurs résultats peuvent être généralisés ou mis en perspective par rapport aux élections contemporaines et futures. Nous présenterons donc le déroulement d'une campagne électorale type liée à l'utilisation des données massives.

Traditionnellement, le parti ou le·la candidat·e au pouvoir bénéficie d'un avantage certain face à ses adversaires [Downs, 1957]. En effet, dans certains pays il est capable de déclencher des élections si cela peut l'avantager, notamment lorsque le parti au pouvoir est majoritaire et que la date des scrutins électoraux n'est pas à date fixe. De plus, lors d'un scrutin, les électeurs vont évaluer le parti/candidat sortant face à ses réalisations tandis que les nouveaux entrants ne seront associés qu'à leurs promesses de campagne ou au souvenir des actions de leur parti s'il fut par le passé au pouvoir. En parallèle, un candidat sortant peut se servir de son temps à la direction du pays pour avancer certaines futures thématiques de campagne potentielles, et donc arriver "préparé" lors d'un prochain scrutin électoral.

Toutefois, on remarque que l'avantage au candidat sortant ne perdure plus. Que ce soit à travers les trois articles de la thèse ou à travers d'autres élections (États-Unis en 2016, France en 2017, Italie en 2017 ou Ontario en 2018 pour ne citer que ces régions), les partis au pouvoir ne réussissent pas à rester en position et à former un gouvernement. Au contraire, l'avantage semble être au candidat accaparant l'attention médiatique, que ce soit de manière positive ou négative. Des parallèles peuvent être établis entre l'élection d'Emmanuel Macron en 2017 et l'élection de Donald Trump en 2016 : en effet, l'un mena une campagne jugée positive tandis que le second mena une campagne apparemment négative, même si tous deux furent au centre de la couverture médiatique.

En effet, les deux candidats ont fait face à des politiciens aguerris et établis. Par contre, plusieurs éléments ont caractérisé leurs campagnes électorales. Les efforts de campagne furent canalisés à travers les plateformes d'analyse de données, que ce soit pour la gestion des bénévoles (avec Nation Builder ou Cinquante plus un) ou pour le ciblage de publicité et de messages personnalisés avec Cambridge Analytica. Un autre élément commun aux deux campagnes fut le fait que les candidats polarisèrent l'attention médiatique, les autres partis politiques réagissant aux agissements des challengers.

Ces deux campagnes électorales sont loin d'être anecdotiques. Ce que certains partis politiques nomment "dégagisme" réfère en fait à une profonde remise en cause des pouvoirs établis. L'adoption d'Internet depuis les vingt dernières années et plus récemment des médias sociaux

devient alors un catalyseur d'avis citoyens amplifiant les revendications et demandant plus âprement des comptes aux pouvoirs élus. Une des manifestations les plus frappantes furent les révolutions du printemps arabe de 2011, où une série de gouvernements d'Afrique du Nord et du Moyen-Orient furent renversés ou remis en cause par la volonté populaire [Ghonim, 2012].

En décortiquant le déroulement d'une campagne électorale, il apparaît que la succession de journées menant au vote des citoyens est hautement cruciale. Avant le lancement d'une campagne électorale, la préparation de celle-ci est l'élément primordial de toute formation politique. Les efforts de levée de fonds ou de recrutement des bénévoles doivent être optimisés afin de permettre aux formations de mener à terme la campagne électorale. Si l'on se place dans la situation de l'élection québécoise de 2014, après un an au sein de l'opposition officielle, le Parti Libéral du Québec fut dès le départ prêt à mener sa campagne électorale, tandis que le Parti Québécois ne put canaliser le message présenté aux médias. La méthodologie utilisée par les instituts de sondage permet de présenter une photographie à un temps t de l'état de la campagne électorale. Toutefois, cette photographie ne repose que sur les réponses d'au mieux quelques milliers d'individus. Certes, ces questionnaires sont calibrés pour refléter la diversité de la population et mimer les cohortes d'électeurs. Toutefois, un biais de réponse peut subsister puisque les individus ne répondent pas directement à certaines questions face à des interrogateurs externes. Les données massives issues des médias sociaux, bien que ne reflétant pas parfaitement les populations, procurent une occasion idéale d'obtenir en temps réel une nouvelle source de données. Au lieu d'une photographie prise tous les trois jours, une rétroaction peut être réalisée à la minute près, notamment lors de débats télévisés, et permettre ainsi de déterminer les propositions électorales sur lesquelles les citoyens réagissent le plus. C'est donc un flux d'avis continu qui est maintenant accessible aux chercheurs.

Une autre composante essentielle des campagnes électorales est depuis l'élection du président Kennedy aux États-Unis la présence de débats télévisés. Si peu de personnes peuvent se targuer de suivre avec précision l'évolution des thématiques de campagne au jour le jour lors d'une élection, tandis que la proportion d'individus devant leur téléviseur écoutant les candidats est bien plus importante. Les téléphones intelligents apparaissent comme une fenêtre de participation et de réflexion des avis citoyens, où chacun a l'opportunité de réagir en temps réel aux phrases et aux attaques lancées par les candidats. Ce qui pourrait apparaître comme politique-spectacle ne peut être qualifié que de marketing politique, mais permet aux électeurs d'obtenir des indices sur les positions des candidats. L'extension digitale de la participation citoyenne amplifie les messages politiques, où chacun peut rechercher de l'information en temps réel ou échanger avec d'autres individus. Ce dernier élément devient aussi un risque pour la société, puisque la présence de chambres d'écho dans les discussions en ligne permet aux utilisateurs de renforcer le poids de leurs avis personnels [Barberá *et al.*, 2015].

Ces nouvelles campagnes électorales sont donc propices à la persistance de certains évènements dans la psyché populaire. Les scandales de campagne deviennent des moments dont l'entropie croît au fil de la campagne, comme ne pouvant plus être contrôlés par les candidats. C'est ce qui arriva au Parti Conservateur lors des élections canadiennes de 2015 puis au Nouveau Parti Démocrate au milieu de la campagne, au parti Les Républicains lors des élections françaises de 2017 ou encore aux Démocrates en 2016 aux États-Unis. Les journaux et les capsules télévisées reprennent ces événements, les citoyens échangent et réagissent face à ces scandales, puis la boucle de rétroaction s'active, puisque les médias proposent un droit de réponse pour comprendre l'évolution des candidats. À moins d'accaparer l'attention médiatique, tout scandale peut entacher le script bien huilé d'une campagne électorale.

Finalement, la performance des dernières journées d'une campagne est cruciale. Ces derniers jours sont souvent associés à un dernier débat télévisé et mène à des heures où l'attention médiatique atteint son paroxysme. Les derniers déplacements stratégiques des candidats sont réalisés en fonction de leurs calculs de campagne. C'est aussi le moment où les sondages traditionnels affinent leurs prévisions et où les indécis doivent en hypothèse avoir pris leur décision. De ces moments-là peuvent être récoltées des données privilégiées afin de faire remonter l'opinion des citoyens. Cela est déjà une pratique courante des partis politiques, puisqu'ils utilisent des listes d'appels afin de faire sortir le vote en fonction des comtés électoraux décisifs. En utilisant des données géolocalisées émises par les individus sur les médias sociaux, des cartographies plus fines des portraits électoraux des régions politiques peuvent être obtenues.

En détaillant le déroulement d'une campagne électorale actuelle, fortement imprégnée et dirigée par les données, il apparaît que les données massives ont des implications à plusieurs niveaux. Afin d'aller plus loin, nous détaillerons les nouvelles perspectives de recherche issues de cette thèse dans la section suivante.

8.3 Vers d'autres perspectives de recherche

Cette thèse doctorale se concentre principalement sur la structuration des données massives dans le cadre démocratique. Les quatre articles de la thèse ont pris pour objet d'étude les élections afin de mettre en évidence les dynamiques électorales sous un autre angle. Toutefois, cette recherche met aussi en relief d'importantes zones d'étude pouvant être abordées. Ainsi, nous détaillerons quatre perspectives de recherche potentielles, soit de nouveaux objets d'études, la question de la surveillance et de la vie privée, l'utilisation de méthodologies adaptées aux sciences sociales et l'apparition de nouveaux modèles de gouvernance.

8.3.1 Modification des objets d'études

Une lacune mise en évidence par la littérature académique est le manque d'études longitudinales concernant l'utilisation des médias sociaux en politique. Cela est probablement dû au fait que les plateformes numériques restent relativement jeunes, Facebook ayant été fondé en 2004 et Twitter l'année suivante. Des projets de recherche à horizon ambitieux doivent voir le jour, présentant une perspective évolutive des pratiques des partis politiques et des citoyens face à l'utilisation des données massives en politique.

En Europe, l'Eurobarometer offre une telle perspective historique par rapport aux avis des citoyens. Il serait nécessaire d'encourager une méthodologie similaire, recueillant les informations issues des médias sociaux avant, pendant et après une période électorale, puis comparer ces informations à celles collectées lors d'une prochaine élection. Dans cette optique, les trois premiers articles de recherche permettent de constituer un socle sur lequel peuvent démarrer plusieurs projets de recherche concernant les usages des médias sociaux au Québec, au Canada et au Nigeria lors des élections. Les prochaines élections dans ces trois régions (respectivement en 2018, 2019 et 2019) seront donc à surveiller pour mettre en perspective les résultats préalablement obtenus.

De même, les études menées dans le cadre des trois premiers articles de la thèse pourraient être élargies aux périodes hors élections. La méthodologie utilisée permettrait de caractériser les conversations politiques sur les médias sociaux pour servir par la suite de point de référence concernant la communication politique en campagne électorale.

Les applications des téléphones intelligents utilisent de plus en plus les données de géolocalisation personnelles. Alors que Twitter retire par défaut la provenance géographique des messages, une avenue de recherche serait de lier les données extraites de Twitter aux données de coordination électorale. En effet, les partis politiques ont des effectifs et des ressources financières limités, mais possèdent des données internes inédites comme les listes d'appels, les dons des individus ou les sympathisants des différents candidats. Combiner ces données internes des partis politiques avec une cartographie des avis des individus lors des élections permettrait de dresser le portrait complet d'une campagne électorale, avec les stratégies partisanes et les perceptions des individus. Pour dépasser le cadre des élections, une étude de la perception des institutions pourrait être effectuée, ajoutant un prisme d'analyse à la vie démocratique.

Ces nouvelles données sont aussi une porte d'entrée vers l'étude de l'activisme citoyen. Les printemps arabes de 2011 [Ghonim, 2012] ou les manifestations étudiantes de Hong Kong de 2014 (appelées Révolution des parapluies) sont une émanation extrême de l'organisation

des citoyens en contexte politique. Les organisations non gouvernementales dirigent l'avis citoyen lors de débats de société, avec par exemple la question de la gestion des ressources pétrolières au Canada. Les stratégies mises en oeuvre par ces organisations pourraient être analysées par le biais des données massives, notamment lors des campagnes d'influence de l'opinion publique. Lors de débats télévisés, une coordination des messages émis sur les médias sociaux peut être repérée et la thématique promue suivie à travers le temps et à travers les différents comptes utilisés.

À l'instar de la fintech qui concerne les technologies utilisées en finance, les mouvements démocratiques possèdent aussi leurs technologies propres regroupées sous l'étiquette de civic-tech. Ces technologies variées promeuvent la participation citoyenne dans la vie politique, le suivi et l'établissement de budgets participatifs jusqu'à la présentation et la co-construction de lois. Toutefois, ces technologies reposent principalement sur l'utilisation d'applications digitales, générant ainsi de fortes quantités de données. Une mutation de l'objet d'étude de cette thèse doctorale serait donc d'incorporer ce type de données pour comprendre les processus d'implication citoyenne autant lors d'élections qu'en périodes non électorales.

8.3.2 Méthodologies de recherche adaptées aux sciences sociales

Les quatre articles de la thèse mettent en avant le besoin de méthodologies adaptées au traitement des données massives. Ces protocoles de recherche doivent refléter les standards de rigueur scientifique (réplicabilité, objectivité, standardisation des méthodologies...). La réplicabilité des résultats est un élément essentiel de toute recherche en science sociale, permettant de vérifier la validité externe des résultats proposés.

Reposant initialement sur des méthodologies développées à l'aide du logiciel Stata, cette thèse a migré vers un langage universel en R. Un tel langage de programmation permet d'assurer l'utilisation de routines et de méthodologies développées par l'ensemble des chercheurs académiques à travers le monde. Issu de la science informatique, le langage de programmation Python est aussi une autre alternative. Plus traditionnellement utilisé par les sciences sociales, R permet le traitement de données massives et la structuration de ces données à travers des techniques économétriques et une visualisation complexe. Toutefois, c'est la disponibilité des données utilisées lors des projets de recherche qui permet aux équipes de chercheurs d'effectuer des percées scientifiques. À titre d'exemple, le Manifesto Project a permis la publication de plusieurs centaines d'articles scientifiques à partir d'une même base de données ayant une granularité poussée. Dans cette optique, les publications académiques encouragent le partage de jeux de données ou la publication de jeux de données inédites, comme fait par exemple dans le cadre du quatrième article de la thèse où les données produites ont été soumises à

publication.

Ces méthodologies de recherche sont donc en évolution, tout comme les pratiques des différents laboratoires de recherche. L'innovation liée à ces méthodologies permet notamment de répondre aux questions traditionnelles du champ de recherche de la science politique, mais aussi de répondre à de nouvelles questions de recherche. Par exemple, comment se structure la participation citoyenne lors d'une campagne électorale ? Les données relatives à ce genre de questionnement sont issues de sondages ou de prises de données à la sortie des bureaux de vote. Toutefois, il est maintenant possible de cartographier par quartier les déterminants socio démographiques des différentes villes, et donc d'inférer statistiquement les influences potentielles du taux d'abstention. Une telle méthodologie quantitative n'aurait pu être mise en oeuvre sans les avancées de la science des données.

Les médias sociaux restent encore un médium jeune dont le travail de structuration est en évolution. Les nombreux articles scientifiques publiés démontrent l'effervescence lié à ce domaine de recherche, et les innovations technologiques des géants du web transforment les habitudes d'utilisation des individus. Le champ de recherche liée à l'utilisation des données massives lors des élections sera donc à l'avenir un champ de recherche en changement, et les données obtenues pourront être employées à travers les différents champs disciplinaires.

8.3.3 Question de la surveillance, de vie privée et des fausses nouvelles

Les élections américaines de 2016 et la campagne électorale du Brexit ont révélé la présence de contenus automatisés sur les plateformes sociales en ligne. Ces robots, combinés avec le ciblage de contenus auprès d'un ensemble de population de plus en plus spécifique, ont formalisé le concept de propagande algorithmique. Les publicités sur Facebook peuvent être dirigées auprès d'un échantillon d'individus présentant certaines caractéristiques ; les méthodes de marketing des marques étant mises à profit dans le cadre des campagnes électorales. La propagation de fausses nouvelles ("fake news") semble avoir influencé le discours médiatique relié à différents scrutins électoraux. La recherche académique peut désormais aborder ce nouveau terrain de recherche. En utilisant des méthodologies d'analyse de réseaux, il est possible de comprendre la transmission de messages et les conditions permettant à ce qu'un contenu en particulier devienne viral ou influence une portion de l'électorat.

Un des apports de cette thèse doctorale est la construction de nouveaux indicateurs pour suivre le cours d'une élection ou caractériser des phénomènes survenant lors d'élections. L'article 4 se penche sur la question du populisme, terme galvaudé à travers les médias au cours des dernières années. Ainsi, la science des données permet d'aborder cette thématique afin de comprendre et de mesurer les différentes dimensions du populisme à travers les pays.

Toutefois, les questions de surveillance et de vie privée restent centrales par rapport à l'utilisation des données massives en politique. L'implication de telles méthodologies permettrait d'inférer et d'anticiper les actions politiques des individus. De par le rôle central des compagnies privées dans la production et la valorisation monétaire des données personnelles, la légitimité démocratique des plateformes technologiques est à surveiller. Les audiences publiques du fondateur de Facebook auprès du Sénat américain, du Parlement européen et du Parlement britannique sont une illustration des volontés d'imputabilité de ces plateformes, tant leur position apparaît systémique dans la vie quotidienne des individus et des démocraties. Plusieurs pays (notamment sur le continent africain) bloquent encore l'accès à Internet ou aux réseaux sociaux lors des élections.

8.3.4 Vers de nouveaux modèles de gouvernance

Finalement, un dernier axe de recherche qui émerge concerne l'innovation technologique que sont les chaînes de blocs (blockchain). La transparence associée à plusieurs de ces protocoles informatiques permet d'accéder à de nouvelles données et donc de les incorporer dans des projets de recherche originaux. Les applications étant à l'origine en finance, de nombreux projets d'application ont vu le jour depuis, notamment en politique (Flux, DemocracyEarth, LaPrimaire.org), en certification de produits (Walmart et IBM), en suivi de la chaîne d'approvisionnement (Provenance) ou en identification des individus (BlockStack). Développer cet axe de recherche serait une occasion parfaite pour ancrer cette technologie dans le domaine académique.

Les chaînes de blocs reposent sur plusieurs principes, notamment la transparence, l'intermédiarité du réseau et la cryptographie [Nakamoto, 2008]. Les données produites par le biais de transactions entre utilisateurs sont accessibles à tout individu, offrant de nouvelles applications potentielles, notamment dans le cadre démocratique. En effet, comment assurer la légitimité du comptage électoral dans des pays où les urnes peuvent être falsifiées ou dans des pays où les terminaux de votes peuvent être hackés ? Comment assurer l'intégrité du vote à travers Internet ? Comment déléguer le vote d'une personne à une autre ? Ce sont des types de question auxquelles peuvent répondre les applications reposant sur les chaînes de blocs. Une institutionnalisation de ces nouvelles technologies est en cours, notamment aux États-Unis, au Canada et en France à travers les Civic Halls de New York, Toronto et Paris, mais aussi à l'échelle national avec l'Estonie et la dématérialisation des procédures légales (e-Estonia).

CHAPITRE 9 CONCLUSION ET RECOMMANDATIONS

Comment les données massives et la science des données peuvent être utilisées pour comprendre les processus démocratiques à l'ère d'Internet ? Telle fut la question de recherche structurant cette thèse doctorale. Les quatre articles tentèrent d'y répondre en abordant chacun un angle d'approche différent et en utilisant soit des données inédites issues des médias sociaux soit des données trop importantes pour être traitées traditionnellement. Les techniques de science de données ont permis d'utiliser et de structurer ces données massives.

Pour conclure cette thèse, une revue des apports sera donc dressée, ainsi que les limites de la recherche effectuée et les recommandations faisant suite à ce travail.

9.1 Apports méthodologiques, théoriques, thématiques

Au niveau méthodologique, plusieurs contributions ont été effectuées. Tout d'abord, l'entièreté des trois premiers articles de recherche repose sur de nouvelles données issues de Twitter. Les élections de 2014 au Québec, du 2015 au Canada et de 2015 au Nigeria ont ainsi été suivies à partir de l'élaboration de bases de données inédites. À travers les quatre articles de recherche, de nombreux indices caractérisant les élections ou les partis politiques ont été construits avec des techniques de traitement du langage naturel. Une autre contribution méthodologique a été l'adaptation de mesures de similarité textuelles aux manifestes politiques européens. L'originalité de cette méthodologie fut de caractériser les partis politiques malgré la difficulté de comparer des textes ayant été écrits en différentes langues. Toujours au niveau méthodologique, de nombreux modèles économétriques ont été présentés, et plus particulièrement à travers le troisième article de la thèse. De manière générale, ces méthodologies quantitatives issues de la science des données ont permis de structurer les données massives liées au domaine politique.

Au niveau théorique, deux contributions ont été développées. Dans un premier temps, les trois articles de recherche ont illustré la défaite de candidats préalablement en poste. À l'ère des médias sociaux, il reste plus important de maîtriser l'attention médiatique que de compter sur les expériences au pouvoir. Les élections étudiées au Québec, au Canada et au Nigeria en font état, mais de nombreuses autres élections présentent les mêmes caractéristiques (France 2017, États-Unis 2016, Mexique 2018, Italie 2018 pour ne citer que celles-ci). Pour confirmer si l'avantage aux candidats sortants est modifié à une époque marquée par les réseaux sociaux, il faudrait toutefois étendre la portée des trois premiers articles pour considérer l'ensemble

des élections contemporaines. Le second apport théorique est une tentative de caractérisation quantitative de ce que l'on nomme "partis de gouvernement" vis-à-vis des partis dits "populistes". Le quatrième article de la thèse présente un panorama européen des discours de partis politiques et permet l'obtention de nouvelles mesures quant au positionnement des partis politiques.

Finalement, les quatre articles de la thèse présentent des apports thématiques concernant les élections respectives couvertes. Chaque article est une illustration d'un événement récent ayant eu lieu au cours du travail doctoral, tandis que le quatrième article de recherche fait état de l'historique des partis politiques européens depuis le début des années 2000 jusqu'en 2018.

9.2 Limites

Plusieurs limites méthodologiques ont été relevées à travers la littérature académique. Toutefois, il est à noter que de plus en plus d'articles scientifiques utilisent les médias sociaux, et Twitter en particulier, comme source d'informations. La publication instantanée des messages, la fine granularité des données, la diversité des informations contenues dans les métadata, l'interaction entre les utilisateurs et la possibilité de bâtir des bases de données en font une source privilégiée. À travers la thèse, quelques limites peuvent être mentionnées.

Tout d'abord, les trois premiers articles (élections au Québec, au Canada et au Nigeria) utilisent des données provenant de Twitter. Ces données ont été sélectionnées à partir de mots-clics (les trois articles) ou de zones géographiques particulières (article 3). Le désavantage de cette méthode est que l'ensemble de l'information n'est pas considéré. Aussi, le but de ces articles ne fut pas de prédire des élections, mais bien de comprendre et de structurer l'information faisant référence à ces élections. Pour parfaitement analyser les interactions des citoyens lors d'une élection, il est nécessaire de considérer d'autres plateformes numériques mais aussi les individus n'utilisant pas de tels services. Une autre limite méthodologique liée à l'utilisation des données de Twitter concerne la construction d'indices et de thématiques pour le suivi des élections. Dans le cadre de l'article 1, les notions de laïcité et d'indépendance ont été regroupées sous une même thématique. Il serait intéressant pour une prochaine étude concernant les messages publiés sur Twitter dans le cadre des élections québécoises de départager ces deux notions afin de refléter les sujets de société actuels présents dans l'actualité québécoise, qui plus est si une de ces notions devient la question de l'urne lors de l'élection.

Lors du quatrième article de la thèse, seuls 676 manifestes politiques furent analysés. En effet, plusieurs manifestes politiques n'ont pas encore été codés et assemblés dans le Mani-

festo Project. À terme, une étude reprenant l'ensemble des manifestes politiques devrait être envisagée, même si les résultats obtenus jusqu'à présents resteront valides puisque la comparaison entre les partis politiques déjà référencés a été effectuée. Une autre limitation liée à cette base de données provient du fait que seuls les partis ayant remporté un siège de député lors des élections figurent au sein de la base de données ; une telle étude pourrait aussi être étendue à d'autres partis, entre autres moins visibles lors des élections.

Finalement, une des limites de cette recherche est le besoin de généralisation des résultats des différents articles. Chaque suivi d'élection s'ancre dans une littérature existante et grandissante. Toutefois, le fait que les plateformes numériques restent des objets d'étude jeunes devrait demander une attention particulière de la part des chercheurs pour le suivi de prochaines élections ayant lieu sur les mêmes terrains d'observation. Ainsi, les résultats obtenus dans cette thèse (ou à travers les articles de la littérature académique) pourraient être mis en perspective avec justesse.

9.3 Recommandations

Des limites de la recherche peuvent être dressées deux recommandations. Dans un premier temps, il est nécessaire de construire des bases de données longitudinales concernant les messages publiés sur Twitter lors des élections. Ces bases de données contribueront à la généralisation des résultats de recherche sur le sujet. De plus, ces bases de données devraient idéalement être consignées par des instituts de recherche ou des universités de manière ouverte, pour que l'ensemble des chercheurs puisse utiliser des données semblables et ainsi développer une culture de recherche reproductible. Cela éviterait aussi que de telles données ne soient que consignées par les partis politiques à des fins de mobilisation électorale. Au delà des partis politiques, la constitution de ces bases de données permettrait notamment aux chercheurs de s'affranchir des contraintes liées aux conditions d'accès établies par les géants du numérique. Twitter reste un réseau social favorisé pour l'étude de phénomènes communicationnels de par la nature ouverte de son API, mais les conditions d'accès peuvent être modifiées sans préavis. Le passage de la limite de caractères de 140 à 280 pour Twitter ou la volonté récente (2019) de Facebook d'encrypter l'ensemble des données de ses services sont des illustrations du rapport de force inégal entre le monde universitaire et les géants numériques.

Une seconde recommandation est la volonté d'ouvrir les techniques d'analyse présentées dans cette thèse à de nouveaux objets de recherche, notamment d'autres applications utilisées par les citoyens. Instagram est un exemple, étant un service de partage instantané de photographies, de messages et de commentaires. WeChat, plateforme provenant de Chine, compte

plus d'un milliard d'utilisateurs. Les applications développées sous l'appellation des "civic-tech" (technologies à usage civique) sont aussi un objet d'étude particulièrement intéressant. Ces dernières restent récentes, et leurs usages peu exploités à travers la littérature académique. De nouvelles technologies sont au fil des années utilisées en contexte électoral, notamment les chaînes de blocs (blockchain), dont l'étude bénéficierait des méthodes structurantes développées dans cette thèse.

En reprenant les trois questions posées en introduction de cette thèse : comment évolue le paysage démocratique, comment se déroule une élection à l'ère d'Internet, et comment se structurent les interactions entre citoyens et représentants politiques ? Si près de trente ans de démocratisation d'Internet et une quinzaine d'années suivant l'émergence des réseaux sociaux ont modifié le contexte politique de nos sociétés, qu'en sera-t-il à l'orée des années 2040 ? Nous avions débuté l'aventure de cette thèse avec un observateur au pied des ruines du Mur de Berlin, mais que pourra dire l'observateur contemporain situé sur les fondations du mur entre le Mexique et les États-Unis concernant les futures décennies ?

Il est un exercice complexe que d'effectuer de la prospection technologique, puisque les avancées des vingt dernières années ne purent être anticipées. Après tout, avant de rencontrer les différents chefs d'États à travers le monde, Amazon ne fut que revendeur de livres, Netflix livra des DVD par la poste, Facebook fut un moyen de comparer des visages entre étudiants et Alphabet (initialement Google) qu'un simple moteur de référencement de pages web.

Toutefois, de nouvelles dynamiques virent le jour depuis. Les habitudes de consommation de l'information des individus se modifièrent. La donnée comme source de valeur est un concept qui ayant de plus en plus d'importance. C'est un aspect central des plateformes technologiques d'Internet et des futurs mastodontes du capitalisme informationnel. On assiste aussi à l'utilisation à des fins stratégiques de fausses informations, nommée propagande algorithmique (en anglais "weaponized algorithmic propaganda"). Il est donc nécessaire de remettre aux citoyens les mêmes moyens d'organisation et de traitement de l'information afin de contrebalancer de potentielles dérives.

Cela passe nécessairement par une évolution progressive concernant l'éducation des individus par rapport à l'information, à la vie privée et aux méthodes d'utilisation d'Internet de manière générale, et des médias sociaux plus précisément. Les prochaines générations seront probablement au courant de tels phénomènes, comme le démontre l'abandon de plateformes telles que Facebook par les plus jeunes utilisateurs. En parallèle, l'utilisation de techniques d'encryptage et la promotion de logiciels ouverts où chacun pourra vérifier et s'assurer des données produites sera une étape essentielle pour la vie démocratique.

Alors comment seront les deux prochaines décennies politiques ? Un mélange subtil où évolueront ces différents phénomènes, où l'organisation et la mobilisation citoyenne seront concomitantes aux stratégies de propagande de partis politiques, où les représentants élus auront accès aux dernières données disponibles pour prendre des décisions éclairées par les faits, où la délégation de la votation s'effectuera en un clic, où l'avis des citoyens pourra être pris en compte en temps réel à travers une fluctuation de sondages permanents.

Vote par vote, caractère par caractère, bit par bit.

RÉFÉRENCES

[Ai et Norton, 2003] Ai, C. et Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129. http://dx.doi.org/10.1016/S0165-1765(03)00032-6. Récupéré le 2018-07-24 de http://www.sciencedirect.com/science/article/pii/S0165176503000326

[Al-Kabi et Al-Sinjilawi, 2007] Al-Kabi, M. N. et Al-Sinjilawi, S. I. (2007). A comparative study of the efficiency of different measures to classify arabic text. *University of Sharjah Journal of Pure and Applied Sciences*, *4*(2), 13–26.

[Aldrich *et al.*, 2011] Aldrich, J. H., Montgomery, J. M. et Wood, W. (2011). Turnout as a habit. *Political Behavior*, *33*(4), 535–563. http://dx.doi.org/10.1007/s11109-010-9148-3. Récupéré le 2018-07-23 de https://link.springer.com/article/10.1007/s11109-010-9148-3

[Alschner *et al.*, 2017] Alschner, W., Seiermann, J. et Skougarevskiy, D. (2017). The impact of the TPP on trade between member countries : A text-as-data approach.

[Alschner et Skougarevskiy, 2016] Alschner, W. et Skougarevskiy, D. (2016). Mapping the universe of international investment agreements. *Journal of international economic law*, *19*(3), 561–588.

[Alvarez *et al.*, 2006] Alvarez, R. M., Boehmke, F. J. et Nagler, J. (2006). Strategic voting in british elections. *Electoral Studies*, *25*(1), 1–19. http://dx.doi.org/10.1016/j.electstud.2005.02.008. Récupéré le 2018-07-23 de http://www.sciencedirect.com/science/article/pii/S0261379405000260

[Anderson, 1984] Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–30.

[Ansolabehere *et al.*, 1994] Ansolabehere, S., Iyengar, S., Simon, A. et Valentino, N. (1994). Does attack advertising demobilize the electorate ? *American political science review*, *88*(4), 829–838.

[Ansolabehere *et al.*, 2008] Ansolabehere, S., Rodden, J. et Snyder, J. M. (2008). The strength of issues : Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, *102*(2), 215–232. http://dx.doi.org/10.1017/S0003055408080210. Récupéré le 2018-07-23 de https://www.cambridge.org/core/journals/american-political-science-review/article/the-strength-of-issues-using-multiple-measures-to-gauge-preference/

-stability-ideological-constraint-and-issue-voting/16F6AF97F7B71AA0112EC9ADF78B553A

[Ansolabehere *et al.*, 2001] Ansolabehere, S., Snyder, J. M. et Stewart, C. (2001). Candidate positioning in u.s. house elections. *American Journal of Political Science*, *45*(1), 136–159. http://dx.doi.org/10.2307/2669364. Récupéré le 2018-07-23 de http://www.jstor.org/stable/2669364

[Arnold *et al.*, 2018] Arnold, J. B., Daroczi, G., Werth, B., Weitzner, B., Kunst, J., Auguie, B., Rudis, B., package.), H. W. C. f. t. g., package), J. T. C. f. t. l. et London, J. (2018). *ggthemes : Extra Themes, Scales and Geoms for 'ggplot2'*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=ggthemes

[Asuncion *et al.*, 2009] Asuncion, A., Welling, M., Smyth, P. et Teh, Y. W. (2009). On smoothing and inference for topic models. Dans *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 27–34. AUAI Press.

[Ausserhofer et Maireder, 2013] Ausserhofer, J. et Maireder, A. (2013). National politics on twitter. *Information, Communication & Society*, *16*(3), 291–314. http://dx.doi.org/10.1080/1369118X.2012.756050. Récupéré le 2018-07-23 de https://doi.org/10.1080/1369118X.2012.756050

[Barbera, 2018] Barbera, P. (2018). *streamR : Access to Twitter Streaming API via R*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=streamR

[Barberá *et al.*, 2015] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. et Bonneau, R. (2015). Tweeting from left to right , tweeting from left to right : Is online political communication more than an echo chamber ? , is online political communication more than an echo chamber ? *Psychological Science*, *26*(10), 1531–1542. http://dx.doi.org/10.1177/0956797615594620. Récupéré le 2018-07-23 de https://doi.org/10.1177/0956797615594620

[Barberá et Rivero, 2015] Barberá, P. et Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, *33*(6), 712–729.

[Barone *et al.*, 2015] Barone, G., D'Acunto, F. et Narciso, G. (2015). Telecracy : Testing for channels of persuasion. *American Economic Journal : Economic Policy*, *7*(2), 30–60. http://dx.doi.org/10.1257/pol.20130318. Récupéré le 2018-07-23 de https://www.aeaweb.org/articles?id=10.1257/pol.20130318

[Bartels, 2005] Bartels, L. M. (2005). Homer gets a tax cut : Inequality and public policy in the american mind. *Perspectives on Politics*, *3*(1), 15–31. Récupéré le 2018-07-23 de http://www.jstor.org/stable/3688108

[Beauchesne, 2013] Beauchesne, O. H. (2013). La campagne dans la twittosphère. *Les Québécois aux urnes. Les partis, les médias et les citoyens en campagne, Montréal, pum*, p. 123.

[Benoit *et al.*, 2017] Benoit, K., Muhr, D. et Watanabe, K. (2017). *stopwords : Multilingual Stopword Lists*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=stopwords`

[Bermingham et Smeaton, 2011] Bermingham, A. et Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. Récupéré le 2018-07-23 de `http://doras.dcu.ie/16670/`

[Betz, 2004] Betz, H.-G. (2004). The growing threat of the radical right. In *Right-wing extremism in the twenty-first century* 85–104. Routledge.

[Betz et Johnson, 2004] Betz, H.-G. et Johnson, C. (2004). Against the current—stemming the tide : the nostalgic ideology of the contemporary radical populist right. *Journal of Political Ideologies*, *9*(3), 311–327. `http://dx.doi.org/10.1080/1356931042000263546`. Récupéré le 2018-07-25 de `https://doi.org/10.1080/1356931042000263546`

[Blais *et al.*, 2009] Blais, A., Gidengil, E., Fournier, P. et Nevitte, N. (2009). Information, visibility and elections : Why electoral outcomes differ when voters are better informed. *European Journal of Political Research*, *48*(2), 256–280. `http://dx.doi.org/10.1111/j.1475-6765.2008.00835.x`. Récupéré le 2018-07-23 de `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6765.2008.00835.x`

[Blais et Indridason, 2007] Blais, A. et Indridason, I. H. (2007). Making candidates count : The logic of electoral alliances in two-round legislative elections. *Journal of Politics*, *69*(1), 193–205. `http://dx.doi.org/10.1111/j.1468-2508.2007.00504.x`. Récupéré le 2018-07-23 de `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2508.2007.00504.x`

[Blais *et al.*, 2011] Blais, A., Lachat, R., Hino, A. et Doray-demers, P. (2011). The mechanical and psychological effects of electoral systems : A quasi-experimental study. *Comparative Political Studies*, 1599–1621.

[Blei *et al.*, 2003] Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

[Bonacich, 1972] Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, *2*(1), 113–120.

[Brady *et al.*, 1995] Brady, H. E., Verba, S. et Schlozman, K. L. (1995). Beyond ses : A resource model of political participation. *The American Political Science Review*, *89*(2), 271–294. `http://dx.doi.org/10.2307/2082425`. Récupéré le 2018-07-23 de `http://www.jstor.org/stable/2082425`

[Brin et Page, 1998] Brin, S. et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Récupéré le 2018-07-23 de http://ilpubs.stanford.edu:8090/361/

[Bruns et Burgess, 2011] Bruns, A. et Burgess, J. E. (2011). #ausvotes : how twitter covered the 2010 australian federal election. *Communication, Politics and Culture*, *44*, 37–56. Récupéré le 2018-07-23 de http://search.informit.com.au/documentSummary;dn=627330171744964;res=IELHSS

[Bruns et Moe, 2014] Bruns, A. et Moe, H. (2014). Structural layers of communication on twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, et C. Puschmann (dir.), *Twitter and Society*, volume 89 15–28. Peter Lang

[Bruns et Stieglitz, 2013] Bruns, A. et Stieglitz, S. (2013). Towards more systematic twitter analysis : metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91–108.

[Bruns et Himmler, 2011] Bruns, C. et Himmler, O. (2011). Newspaper circulation and local government efficiency. *The Scandinavian Journal of Economics*, *113*(2), 470–492. http://dx.doi.org/10.1111/j.1467-9442.2010.01633.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9442.2010.01633.x

[Budge et Klingemann, 2001] Budge, I. et Klingemann, H.-D. (2001). *Mapping policy preferences : estimates for parties, electors, and governments, 1945-1998*, volume 1. Oxford University Press on Demand.

[Burgess et Bruns, 2012] Burgess, J. et Bruns, A. (2012). (not) the twitter election. *Journalism Practice*, *6*(3), 384–402. http://dx.doi.org/10.1080/17512786.2012.663610. Récupéré le 2018-07-23 de https://doi.org/10.1080/17512786.2012.663610

[Butts, 2016] Butts, C. T. (2016). *sna : Tools for Social Network Analysis*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=sna

[Butts *et al.*, 2018] Butts, C. T., Hunter, D., Handcock, M., Bender-deMoll, S. et Horner, J. (2018). *network : Classes for Relational Data*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=network

[Calvo et Hellwig, 2010] Calvo, E. et Hellwig, T. (2010). Centripetal and centrifugal incentives under different electoral systems. *American Journal of Political Science*, *55*(1), 27–41. http://dx.doi.org/10.1111/j.1540-5907.2010.00482.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2010.00482.x

[Cameron *et al.*, 2013] Cameron, M. P., Barrett, P. et Stewardson, B. (2013). Can social media predict election results ? evidence from new zealand. Récupéré le 2018-07-23 de https://ideas.repec.org/p/wai/econwp/13-08.html

[Campante *et al.*, 2013] Campante, F. R., Durante, R. et Sobbrio, F. (2013). Politics 2.0 : The multifaceted effect of broadband internet on political participation. http://dx.doi.org/10.3386/w19029. Récupéré le 2018-07-23 de http://www.nber.org/papers/w19029

[Carsey et Layman, 2006] Carsey, T. M. et Layman, G. C. (2006). Changing sides or changing minds ? party identification and policy preferences in the american electorate. *American Journal of Political Science*, *50*(2), 464–477. http://dx.doi.org/10.1111/j.1540-5907.2006.00196.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2006.00196.x

[Castillo *et al.*, 2011] Castillo, C., Mendoza, M. et Poblete, B. (2011). Information credibility on twitter. Dans *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, 675–684. ACM. http://dx.doi.org/10.1145/1963405.1963500. Récupéré le 2018-07-23 de http://doi.acm.org/10.1145/1963405.1963500

[Cha *et al.*, 2012] Cha, M., Benevenuto, F., Haddadi, H. et Gummadi, K. (2012). The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans*, *42*(4), 991–998.

[Chong et Druckman, 2007] Chong, D. et Druckman, J. N. (2007). Framing public opinion in competitive democracies. *American Political Science Review*, *101*(4), 637–655. http://dx.doi.org/10.1017/S0003055407070554. Récupéré le 2018-07-23 de https://www.cambridge.org/core/journals/american-political-science-review/article/framing-public-opinion-in-competitive-democracies/BCB2AC51F4AC3623ED9E6BEDF7DD854E

[Choy *et al.*, 2012] Choy, M., Cheong, M., Ma, N. L. et Koo, P. S. (2012). A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *Research Collection School Of Information Systems*. Récupéré de https://ink.library.smu.edu.sg/sis_research/1436

[Chung et Mustafaraj, 2011] Chung, J. et Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections ? volume 11. Récupéré de https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3549/4126

[Cogburn et Espinoza-Vasquez, 2011] Cogburn, D. L. et Espinoza-Vasquez, F. K. (2011). From networked nominee to networked nation : Examining the impact of web 2.0 and social media on political participation and civic engagement in the 2008 obama campaign. *Journal of Political Marketing*, *10*(1), 189–213. http://dx.doi.org/10.1080/15377857.2011.540224. Récupéré le 2018-07-23 de https://doi.org/10.1080/15377857.2011.540224

[Collingwood *et al.*, 2013] Collingwood, L., Jurka, T., Boydstun, A. E., Grossman, E. et van Atteveldt, W. H. (2013). RTextTools : A supervised learning package for text classification.

[Conover *et al.*, 2011] Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A. et Menczer, F. (2011). Predicting the political alignment of twitter users. 192–199. http://dx.doi.org/10.1109/PASSAT/SocialCom.2011.34

[Couture-Beil, 2018] Couture-Beil, A. (2018). *rjson : JSON for R.* Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=rjson

[Csardi et al., 2018] Csardi, G. et al. (2018). *igraph : Network Analysis and Visualization.* Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=igraph

[Csardi et Nepusz, 2006] Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

[de Marcellis-Warin *et al.*, 2015] de Marcellis-Warin, N., Sanger, W. et Warin, T. (2015). Assessing influence on social media : reputation risks in networks. Dans *ECSM2015-Proceedings of the 2nd European Conference on Social Media 2015 : ECSM 2015*, volume 313. Academic Conferences Limited.

[DellaVigna et Kaplan, 2007] DellaVigna, S. et Kaplan, E. (2007). The fox news effect : Media bias and voting. *The Quarterly Journal of Economics*, *122*(3), 1187–1234. http://dx.doi.org/10.1162/qjec.122.3.1187. Récupéré le 2018-07-23 de https://academic.oup.com/qje/article/122/3/1187/1879517

[DiMaggio *et al.*, 2013] DiMaggio, P., Nag, M. et Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture : Application to newspaper coverage of u.s. government arts funding. *Poetics*, *41*(6), 570–606. http://dx.doi.org/10.1016/j.poetic.2013.08.004. Récupéré le 2018-07-24 de http://www.sciencedirect.com/science/article/pii/S0304422X13000661

[Dow et Endersby, 2004] Dow, J. K. et Endersby, J. W. (2004). Multinomial probit and multinomial logit : a comparison of choice models for voting research. *Electoral Studies*, *23*(1), 107–122. http://dx.doi.org/10.1016/S0261-3794(03)00040-4. Récupéré le 2018-07-24 de http://www.sciencedirect.com/science/article/pii/S0261379403000404

[Downs, 1957] Downs, A. (1957). *An economic theory of democracy.* Harper. OCLC : 254197.

[Drakos et Kouretas, 2015] Drakos, A. A. et Kouretas, G. P. (2015). The conduct of monetary policy in the eurozone before and after the financial crisis. *Economic Modelling*, *48*, 83–92.

[Dumitrica, 2014] Dumitrica, D. (2014). Politics as "customer relations" : Social media and political authenticity in the 2010 municipal elections in calgary, canada. *Javnost-the public*, *21*(1), 53–69.

[Dupont et Dupont, 2009] Dupont, W. D. et Dupont, W. D. (2009). *Statistical modeling for biomedical researchers : a simple introduction to the analysis of complex data.* Cambridge University Press.

[Durante et Knight, 2012] Durante, R. et Knight, B. (2012). Partisan control, media bias, and viewer responses : Evidence from berlusconi's italy. *Journal of the European Economic Association*, *10*(3), 451–481. http://dx.doi.org/10.1111/j.1542-4774.2011.01060.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1542-4774.2011.01060.x

[Eatwell, 2000] Eatwell, R. (2000). The rebirth of the'extreme right' in western europe ? *Parliamentary affairs*, *53*(3), 407–425.

[Elff, 2013] Elff, M. (2013). A dynamic state-space model of coded political texts. *Political Analysis*, *21*(2), 217–232. http://dx.doi.org/10.1093/pan/mps042. Récupéré le 2018-07-23 de https://www.cambridge.org/core/journals/political-analysis/article/dynamic-statespace-model-of-coded-political-texts/B29D9DD75CCBC37248CCBAF708D24917

[Elmer, 2013] Elmer, G. (2013). Live research : Twittering an election debate , live research : Twittering an election debate. *New Media & Society*, *15*(1), 18–30. http://dx.doi.org/10.1177/1461444812457328. Récupéré le 2018-07-23 de https://doi.org/10.1177/1461444812457328

[Enikolopov *et al.*, 2011] Enikolopov, R., Petrova, M. et Zhuravskaya, E. (2011). Media and political persuasion : Evidence from russia. *American Economic Review*, *101*(7), 3253–3285. http://dx.doi.org/10.1257/aer.101.7.3253. Récupéré le 2018-07-23 de https://www.aeaweb.org/articles?id=10.1257/aer.101.7.3253

[Enli et Naper, 2016] Enli, G. et Naper, A. A. (2016). Social media incumbent advantage : Barack obama's and mitt romney's tweets in the 2012 US presidential election campaign.

[Eyries et Poirier, 2013] Eyries, A. et Poirier, C. (2013). Une communication politique 2.0. approche comparative des usages électoraux de twitter en france et au québec. *Communication. Information médias théories pratiques*, *32*(2).

[Ezrow *et al.*, 2011] Ezrow, L., De Vries, C., Steenbergen, M. et Edwards, E. (2011). Mean voter representation and partisan constituency representation : Do parties respond to the mean voter position or to their supporters ? , mean voter representation and partisan constituency representation : Do parties respond to the mean voter position or to their supporters ? *Party Politics*, *17*(3), 275–301. http://dx.doi.org/10.1177/1354068810372100. Récupéré le 2018-07-23 de https://doi.org/10.1177/1354068810372100

[Feinerer *et al.*, 2018] Feinerer, I., Hornik, K., Software, A. et Ghostscript), I. p. p. t. f. G. (2018). *tm : Text Mining Package*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=tm

[Fellows, 2014] Fellows, I. (2014). *wordcloud : Word Clouds*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=wordcloud

[Filho *et al.*, 2015] Filho, R. M., Almeida, J. M. et Pappa, G. L. (2015). Twitter population sample bias and its impact on predictive outcomes : A case study on elections. Dans *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1254–1261. http://dx.doi.org/10.1145/2808797.2809328

[Fink *et al.*, 2012] Fink, C., Kopecky, J., Bos, N. et Thomas, M. (2012). Mapping the twitterverse in the developing world : An analysis of social media use in nigeria. Dans *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 164–171. Springer.

[Fink *et al.*, 2016] Fink, C., Schmidt, A., Barash, V., Cameron, C. et Macy, M. (2016). Complex contagions and the diffusion of popular twitter hashtags in nigeria. *Social Network Analysis and Mining*, *6*(1), 1.

[Flanagin et Metzger, 2000] Flanagin, A. J. et Metzger, M. J. (2000). Perceptions of internet information credibility , perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, *77*(3), 515–540. http://dx.doi.org/10.1177/107769900007700304. Récupéré le 2018-07-23 de https://doi.org/10.1177/107769900007700304

[Fligstein *et al.*, 2014] Fligstein, N., Brundage, J. et Schultz, M. (2014). Why the federal reserve failed to see the financial crisis of 2008 : The role of "macroeconomics" as a sense making and cultural frame. Récupéré le 2018-07-24 de http://irle.berkeley.edu/why-the-federal-reserve-failed-to-see-the-financial-crisis-of-2008/-the-role-of/macroeconomics-as-a-sense-making-and-cultural-frame/

[Freeman, 1977] Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

[Fruchterman et Reingold, 1991] Fruchterman, T. M. et Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software : Practice and experience*, *21*(11), 1129–1164.

[Gayo-Avello, 2012a] Gayo-Avello, D. (2012a). "i wanted to predict elections with twitter and all i got was this lousy paper" – a balanced survey on election prediction using twitter data. Récupéré de https://arxiv.org/abs/1204.6441

[Gayo-Avello, 2012b] Gayo-Avello, D. (2012b). No, you cannot predict elections with twitter. *IEEE Internet Computing*, *16*(6), 91–94. http://dx.doi.org/10.1109/MIC.2012.137

[Gayo-Avello, 2013] Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, *31*(6), 649–679.

[Gayo Avello *et al.*, 2011] Gayo Avello, D., Metaxas, P. T. et Mustafaraj, E. (2011). Limits of electoral predictions using twitter. Dans *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

[Gentry, 2015] Gentry, J. (2015). *twitteR : R Based Twitter Client*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=twitteR`

[Gentry et Lang, 2015] Gentry, J. et Lang, D. T. (2015). *ROAuth : R Interface For OAuth*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=ROAuth`

[Gentzkow, 2006] Gentzkow, M. (2006). Television and voter turnout. *The Quarterly Journal of Economics*, *121*(3), 931–972. Récupéré le 2018-07-23 de `https://econpapers.repec.org/article/oupqjecon/v_3a121_3ay_3a2006_3ai_3a3_3ap_3a931-972..htm`

[Gerber *et al.*, 2008] Gerber, A. S., Green, D. P. et Larimer, C. W. (2008). Social pressure and voter turnout : Evidence from a large-scale field experiment. *American Political Science Review*, *102*(1), 33–48. `http://dx.doi.org/10.1017/S000305540808009X`. Récupéré le 2018-07-23 de `https://www.cambridge.org/core/journals/american-political-science-review/article/social-pressure-and-voter-turnout-evidence-from-a-largescale/-field-experiment/11E84AF4C0B7FBD1D20C855972C2C3EB`

[Ghonim, 2012] Ghonim, W. (2012). *Revolution 2.0. The Power Of The People Is Greater Than The People In Power : A Memoir*. Houghton Mifflin Harcourt.

[Giasson *et al.*, 2013] Giasson, T., Le Bars, G., Bastien, F. et Verville, M. (2013). L'usage du web social par les partis politiques au québec. le cas de# qc2012. Dans *annual conference of the Canadian Political Science Association, Victoria, Canada*.

[Gilens, 2005] Gilens, M. (2005). Inequality and democratic responsiveness. *Public Opinion Quarterly*, *69*(5), 778–796. `http://dx.doi.org/10.1093/poq/nfi058`. Récupéré le 2018-07-23 de `https://academic.oup.com/poq/article/69/5/778/1920084`

[Glasgow, 2001] Glasgow, G. (2001). Mixed logit models for multiparty elections. *Political Analysis*, *9*(2), 116–136. `http://dx.doi.org/10.1093/oxfordjournals.pan.a004867`. Récupéré le 2018-07-24 de `https://www.cambridge.org/core/journals/political-analysis/article/mixed-logit-models-for-multiparty-elections/00D7056BF23F083D2DC45B285F41388F`

[Glavaš *et al.*, 2017] Glavaš, G., Nanni, F. et Ponzetto, S. P. (2017). Cross-lingual classification of topics in political texts. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 42–46. `http://dx.doi.org/10.18653/v1/W17-2906`. Récupéré le 2018-07-24 de `https://aclanthology.info/papers/W17-2906/w17-2906`

[Golder, 2016] Golder, M. (2016). Far right parties in europe. *Annual Review of Political Science*, *19*, 477–497.

[Golder et Stramski, 2009] Golder, M. et Stramski, J. (2009). Ideological congruence and electoral institutions. *American Journal of Political Science*, *54*(1), 90–106. http://dx.doi.org/10.1111/j.1540-5907.2009.00420.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00420.x

[Gould, 2000] Gould, W. (2000). Interpreting logistic regression in all its forms. *STATA technical bulletin*, *9*(53).

[Graeber, 2011] Graeber, D. (2011). *Debt : The First 5,000 Years*. Melville House.

[Green et Gerber, 2015] Green, D. P. et Gerber, A. S. (2015). *Get out the vote : How to increase voter turnout*. Brookings Institution Press.

[Greene, 2010] Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters*, *107*(2), 291–296. http://dx.doi.org/10.1016/j.econlet.2010.02.014. Récupéré le 2018-07-24 de http://www.sciencedirect.com/science/article/pii/S0165176510000777

[Greene, 2012] Greene, W. (2012). *Econometric Analysis*. Pearson.

[Greenland, 1985] Greenland, S. (1985). An application of logistic models to the analysis of ordinal responses. *Biometrical Journal*, *27*(2), 189–197.

[Grimmer, 2010] Grimmer, J. (2010). A bayesian hierarchical topic model for political texts : Measuring expressed agendas in senate press releases. *Political Analysis*, *18*(1), 1–35.

[Grimmer et Stewart, 2013] Grimmer, J. et Stewart, B. M. (2013). Text as data : The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

[Gruzd et Roy, 2014] Gruzd, A. et Roy, J. (2014). Investigating political polarization on twitter : A canadian perspective. *Policy & Internet*, *6*(1), 28–45. http://dx.doi.org/10.1002/1944-2866.POI354. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1002/1944-2866.POI354

[Grétas *et al.*, 2014] Grétas, C., de Nicolini, et Cimon-Mattar, N. (2014). The national assembly of québec in the digital era. *Canadian Parliamentary Review*, p. 31.

[Grün et Hornik, 2017] Grün, B. et Hornik, K. (2017). *topicmodels : Topic Models*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=topicmodels

[Gschwend, 2006] Gschwend, T. (2006). Ticket-splitting and strategic voting under mixed electoral rules : Evidence from germany. *European Journal of Political Research*, *46*(1), 1–23. http://dx.doi.org/10.1111/j.1475-6765.2006.00641.x. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6765.2006.00641.x

[Handcock *et al.*, 2016] Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S. et Morris, M. (2016). *statnet : Software Tools for the Statistical Analysis of Network Data*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=statnet`

[Handcock *et al.*, 2008] Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. et Morris, M. (2008). statnet : Software tools for the representation, visualization, analysis and simulation of network data. *Journal of statistical software*, *24*(1), 1548.

[Hecht et Stephens, 2014] Hecht, B. et Stephens, M. (2014). A tale of cities : Urban biases in volunteered geographic information. Dans *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 197–205. The AAAI Press. Récupéré le 2018-07-23 de `https://www.scholars.northwestern.edu/en/publications/a-tale-of-cities-urban-biases-in-volunteered-geographic-informati`

[Hilbe, 2009] Hilbe, J. M. (2009). Logistic regression models. *Chapman and Hall/CRC*.

[Holmberg et Oscarsson, 2013] Holmberg, S. et Oscarsson, H. (2013). *Party Leader Effects on the Vote*. Oxford University Press. Récupéré le 2018-07-23 de `http://www.oxfordscholarship.com/view/10.1093/acprof:osobl/9780199259007.001.0001/acprof-9780199259007-chapter-3`

[Hopkins et Ladd, 2014] Hopkins, D. J. et Ladd, J. M. (2014). The consequences of broader media choice : Evidence from the expansion of fox news. *Quarterly Journal of Political Science*, *9*(1), 115–135. `http://dx.doi.org/10.1561/100.00012099`. Récupéré le 2018-07-23 de `https://www.nowpublishers.com/article/Details/QJPS-12099`

[Hornik et Grün, 2011] Hornik, K. et Grün, B. (2011). topicmodels : An r package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

[Hosmer Jr *et al.*, 2013] Hosmer Jr, D. W., Lemeshow, S. et Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

[Hu et Liu, 2004] Hu, M. et Liu, B. (2004). Mining and summarizing customer reviews. Dans *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

[Ifukor, 2010] Ifukor, P. (2010). "elections" or "selections"? blogging and twittering the nigerian 2007 general elections. *Bulletin of Science, Technology & Society*, *30*(6), 398–414.

[Jaccard, 1900] Jaccard, P. (1900). Contribution au problème de l'immigration post-glaciaire de la flore alpine : étude comparative de la flore alpine du massif de wildhorn, du haut bassin du trient et de la haute vallée de bagnes. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *36*, 87–130. `http://dx.doi.org/10.5169/seals-266069`

[Jaccard, 1901a] Jaccard, P. (1901a). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *37*, 241–72. http://dx.doi.org/10.5169/seals-266440

[Jaccard, 1901b] Jaccard, P. (1901b). Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, *37*, 547–579. http://dx.doi.org/10.5169/seals-266450

[Jaccard, 1902a] Jaccard, P. (1902a). Distribution comparée de la flore alpine dans quelques régions des alpes occidentales et orientales. *Bulletin de la Murithienne*, (31), 81–92.

[Jaccard, 1902b] Jaccard, P. (1902b). Lois de distribution florale dans la zone alpine. *Bulletin de la Société vaudoise des sciences naturelles*, *38*, 69–130. http://dx.doi.org/10.5169/seals-266762

[Jaeger, 2008] Jaeger, T. F. (2008). Categorical data analysis : Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007. Récupéré le 2018-07-24 de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613284/

[Jankowski et Gross, 2017] Jankowski, M. et Gross, M. (2017). Analyzing local party manifestos in multi-level democracies. Récupéré de http://www.mzes.uni-mannheim.de/d7/en/publications/presentation/analyzing-local-party-manifestos-in-multi-level-democracies-0

[Johnston *et al.*, 2004] Johnston, R., Hagen, M. G. et Jamieson, K. H. (2004). *The 2000 Presidential Election and the Foundations of Party Politics*. Cambridge University Press. Google-Books-ID : HNqu79LXIqkC.

[Jungherr, 2016] Jungherr, A. (2016). Twitter use in election campaigns : A systematic literature review. *Journal of information technology & politics*, *13*(1), 72–91.

[Jungherr *et al.*, 2012] Jungherr, A., Jürgens, P. et Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions : A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter : What 140 characters reveal about political sentiment". *Social science computer review*, *30*(2), 229–234.

[Jurka *et al.*, 2014] Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E. et Atteveldt, W. v. (2014). *RTextTools : Automatic Text Classification via Supervised Learning*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=RTextTools

[Kalsnes *et al.*, 2014] Kalsnes, B., Krumsvik, A. H. et Storsul, T. (2014). Social media as a political backchannel : Twitter use during televised election debates in norway. *Aslib Journal of Information Management*, *66*(3), 313–328.

[Kamada et Kawai, 1989] Kamada, T. et Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, *31*(1), 7–15.

[Karlsen, 2011] Karlsen, R. (2011). A platform for individualized campaigning ? social media and parliamentary candidates in the 2009 norwegian election campaign. *Policy & Internet*, *3*(4), 1–25. http://dx.doi.org/10.2202/1944-2866.1137. Récupéré le 2018-07-23 de https://onlinelibrary.wiley.com/doi/abs/10.2202/1944-2866.1137

[Kayser et Peress, 2012] Kayser, M. A. et Peress, M. (2012). Benchmarking across borders : Electoral accountability and the necessity of comparison. *American Political Science Review*, *106*(3), 661–684. http://dx.doi.org/10.1017/S0003055412000275. Récupéré le 2018-07-23 de https://www.cambridge.org/core/journals/american-political-science-review/article/benchmarking-across-borders-electoral-accountability-and-the-necessity-/of-comparison/3DC4E5DC7D72FBF7317BBFA117B1AF2D

[Kearney, 2018a] Kearney, M. (2018a). We put data science to the test to try to uncover the mystery author of the times' op-ed. *Reynolds Journalism Institute Stories, https ://www.rjionline.org/stories/we-put-data-science-to-the-test-to-try-to-uncover-the-mystery-author-of-the*.

[Kearney, 2018b] Kearney, M. W. (2018b). *rtweet : Collecting Twitter Data*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=rtweet

[Kendall *et al.*, 2015] Kendall, C., Nannicini, T. et Trebbi, F. (2015). How do voters respond to information ? evidence from a randomized campaign. *American Economic Review*, *105*(1), 322–353. http://dx.doi.org/10.1257/aer.20131063. Récupéré le 2018-07-24 de http://pubs.aeaweb.org/doi/10.1257/aer.20131063

[King, 2011] King, G. (2011). Ensuring the data-rich future of the social sciences. *Science (New York, N.Y.)*, *331*, 719–21. http://dx.doi.org/10.1126/science.1197872

[King et Lowe, 2003] King, G. et Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders : A rare events evaluation design. *International Organization*, *57*, 617–642.

[Kleinbaum et Klein, 2010] Kleinbaum, D. G. et Klein, M. (2010). *Logistic regression : a self-learning text*. Springer Science & Business Media.

[Klingemann *et al.*, 2006] Klingemann, H.-D., Volkens, A., McDonald, M. D., Budge, I. et Bara, J. (2006). *Mapping policy preferences II : estimates for parties, electors, and governments in Eastern Europe, European Union, and OECD 1990-2003*, volume 2. Oxford University Press on Demand.

[Kolaczyk et Csárdi, 2014] Kolaczyk, E. D. et Csárdi, G. (2014). *Statistical analysis of network data with R*, volume 65. Springer.

[Kreiss et McGregor, 2018] Kreiss, D. et McGregor, S. C. (2018). Technology firms shape political communication : The work of microsoft, facebook, twitter, and google with campaigns during the 2016 US presidential cycle. *Political Communication*, *35*(2), 155–177.

[Laney, 2001] Laney, D. (2001). 3-d data management : Controlling data volume, velocity, and variety. *META Group Res Note 6*, *6*.

[Lang et al., 2018] Lang, D. T. et al. (2018). *RCurl : General Network (HTTP/FTP/...) Client Interface for R*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=RCurl`

[Larsson et Moe, 2013] Larsson, A. O. et Moe, H. (2013). Representation or participation ? *Javnost - The Public*, *20*(1), 71–88. `http://dx.doi.org/10.1080/13183222.2013.11009109`. Récupéré le 2018-07-23 de `https://doi.org/10.1080/13183222.2013.11009109`

[Lau et Redlawsk, 2001] Lau, R. R. et Redlawsk, D. P. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*, *45*(4), 951–971. `http://dx.doi.org/10.2307/2669334`. Récupéré le 2018-07-23 de `http://www.jstor.org/stable/2669334`

[Laver *et al.*, 2003] Laver, M., Benoit, K. et Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

[Laver et Garry, 2000] Laver, M. et Garry, J. (2000). Estimating policy positions from political texts. Dans *American Journal of Political Science*, volume 44, 619–634. Wiley-Blackwell. Récupéré le 2018-07-23 de `https://pure.qub.ac.uk/portal/en/publications/estimating-policy-positions-from-political-texts/(a0eece7f-05d8-4f74-88c1-664eb450dd57)/export.html`

[Lehmann *et al.*, 2018] Lehmann, P., Matthiess, T., Merz, N., Regel, S. et Werner, A. (2018). Manifesto corpus. version : 2018-1. Récupéré de `https://manifesto-project.wzb.eu`

[Lewandowski *et al.*, 2018] Lewandowski, J., Merz, N., Regel, S., Lehmann, P. et Muscat, P. (2018). *manifestoR : Access and Process Data and Documents of the Manifesto Project*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=manifestoR`

[Lips, 2014] Lips, M. (2014). *Transforming Government—by Default ?* Oxford University Press. Récupéré le 2018-07-23 de `http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199661992.001.0001/acprof-9780199661992-chapter-12`

[Livne *et al.*, 2011] Livne, A., Simmons, M., Adar, E. et Adamic, L. (2011). The party is over here : Structure and content in the 2010 election. Dans *Fifth International AAAI Conference on Weblogs and Social Media*. Récupéré le 2018-07-23 de `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2852`

[Liégey *et al.*, 2013] Liégey, G., Muller, A. et Pons, V. (2013). *Porte à porte : reconquérir la démocratie sur le terrain*. Calmann-Lévy.

[Lodhi *et al.*, 2002] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. et Watkins, C. (2002). Text classification using string kernels. *J. Mach. Learn. Res.*, *2*, 419–444. `http://dx.doi.org/10.1162/153244302760200687`. Récupéré le 2018-07-24 de `https://doi.org/10.1162/153244302760200687`

[Long *et al.*, 2006] Long, S. J., Long, J. S. et Freese, J. (2006). *Regression models for categorical dependent variables using Stata*. Stata press.

[Lui *et al.*, 2011] Lui, C., Metaxas, P. T. et Mustafaraj, E. (2011). On the predictability of the US elections through search volume activity.

[Lupia, 1994] Lupia, A. (1994). Shortcuts versus encyclopedias : Information and voting behavior in california insurance reform elections. *The American Political Science Review*, *88*(1), 63–76. `http://dx.doi.org/10.2307/2944882`. Récupéré le 2018-07-23 de `http://www.jstor.org/stable/2944882`

[Luskin *et al.*, 2002] Luskin, R. C., Fishkin, J. S. et Jowell, R. (2002). Considered opinions : Deliberative polling in britain. *British Journal of Political Science*, *32*(3), 455–487. `http://dx.doi.org/10.1017/S0007123402000194`. Récupéré le 2018-07-23 de `https://www.cambridge.org/core/journals/british-journal-of-political-science/article/considered-opinions-deliberative-polling-in-britain/9CA53EBEDC6A0C5A39F92A9D991DD4A3`

[Machiavel, 1532] Machiavel, N. (1532). *Le Prince*.

[Malik *et al.*, 2015] Malik, M. M., Lamba, H., Nakos, C. et Pfeffer, J. (2015). Population bias in geotagged tweets. Dans *Ninth International AAAI Conference on Web and Social Media*. Récupéré le 2018-07-23 de `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662`

[Marcellis-Warin *et al.*, 2017] Marcellis-Warin, N. D., Sanger, W. et Warin, T. (2017). A network analysis of financial conversations on twitter. *International Journal of Web Based Communities*, *13*(3), 281–310.

[Mazzoleni *et al.*, 2003] Mazzoleni, G., Stewart, J. et Horsfield, B. (2003). *The Media and Neo-populism : A Contemporary Comparative Analysis*. Greenwood Publishing Group. Google-Books-ID : YdG5cLc_Pi4C.

[Mellon et Prosser, 2017] Mellon, J. et Prosser, C. (2017). Twitter and facebook are not representative of the general population : Political attitudes and demographics of british social media users. *Research & Politics*, *4*(3), 2053168017720008. http://dx.doi.org/10.1177/2053168017720008. Récupéré le 2018-07-23 de https://doi.org/10.1177/2053168017720008

[Metaxas *et al.*, 2011] Metaxas, P. T., Mustafaraj, E. et Gayo-Avello, D. (2011). How (not) to predict elections. Dans *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 165–171. IEEE.

[Mislove *et al.*, 2011] Mislove, A., Jørgensen, S. L., Ahn, Y.-Y., Onnela, J.-P. et Rosenquist, J. N. (2011). Understanding the demographics of twitter users. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 554–557.

[Misra, 2018] Misra, K. (2018). Resistance inside trump's white house - an analysis of authorship. *Github, https ://github.com/kanishkamisra/inside-trumps-white-house*.

[Monroe *et al.*, 2008] Monroe, B. L., Colaresi, M. P. et Quinn, K. M. (2008). Fightin'words : Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.

[Morstatter *et al.*, 2013] Morstatter, F., Pfeffer, J., Liu, H. et Carley, K. M. (2013). Is the sample good enough ? comparing data from twitter's streaming API with twitter's firehose. Dans *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, 400–408. AAAI press. Récupéré le 2018-07-23 de https://asu.pure.elsevier.com/en/publications/is-the-sample-good-enough-comparing-data-from-twitters-streaming-

[Mowbray, 2014] Mowbray, M. (2014). Automated twitter accounts. *Twitter and Society*, 183–194. Récupéré le 2018-07-23 de https://research-information.bristol.ac.uk/en/publications/automated-twitter-accounts(14fd5ab6-a1a7-4849-8e42-97d4909e3087)/export.html

[Mudde, 2000] Mudde, C. (2000). *The Ideology of the Extreme Right*. Manchester University Press.

[Mudde, 2004] Mudde, C. (2004). The populist zeitgeist. *Government and opposition*, *39*(4), 541–563.

[Mudde, 2010] Mudde, C. (2010). The populist radical right : A pathological normalcy. *West European Politics*, *33*(6), 1167–1186.

[Mullen, 2016] Mullen, L. (2016). *textreuse : Detect Text Reuse and Document Similarity*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=textreuse`

[Nakamoto, 2008] Nakamoto, S. (2008). Bitcoin : A peer-to-peer electronic cash system. Récupéré de `http://bitcoin.org/bitcoin.pdf`

[Nickerson, 2008] Nickerson, D. W. (2008). Is voting contagious ? evidence from two field experiments. *American political Science review*, *102*(1), 49–57.

[Nielsen, 2011] Nielsen, F. A. (2011). A new ANEW : Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv :1103.2903*.

[Nieminen, 1974] Nieminen, J. (1974). On the centrality in a graph. *Scandinavian journal of psychology*, *15*(1), 332–336.

[Niwattanakul *et al.*, 2013] Niwattanakul, S., Singthongchai, J., Naenudorn, E. et Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. Dans *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1.

[Odeyemi et Mosunmola, 2015] Odeyemi, T. I. et Mosunmola, O. O. (2015). Stakeholders, ICTs platforms and the 2015 general elections in nigeria. Dans *National Conference, Abudja, Nigeria : The Electoral Institute. Retrieved from http ://www. academia. edu/download/38426835/Conference-Paper-by-Odeyemi-Mosunmola. pdf*.

[Olalekan, 2015] Olalekan, A. (2015). Role of social media in the formation of government policies in nigeria. *The Online Journal of Communication and Media*.

[Onapajo, 2015] Onapajo, H. (2015). How credible were the nigerian 2015 general elections ? *African renaissance*, *12*(3), 11–39.

[Pampel, 2000] Pampel, F. C. (2000). *Logistic regression : A primer*, volume 132. Sage.

[Paradis *et al.*, 2018] Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R., Didier, G., Durand, B., Dutheil, J., Ewing, R. J., Gascuel, O., Guillerme, T., Heibl, C., Ives, A., Jones, B., Krah, F., Lawson, D., Lefort, V., Legendre, P., Lemon, J., McCloskey, R., Nylander, J., Opgen-Rhein, R., Popescu, A.-A., Royer-Carenzi, M., Schliep, K., Strimmer, K. et Vienne, D. d. (2018). *ape : Analyses of Phylogenetics and Evolution*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=ape`

[Paradis *et al.*, 2004] Paradis, E., Claude, J. et Strimmer, K. (2004). APE : analyses of phylogenetics and evolution in r language. *Bioinformatics*, *20*(2), 289–290.

[Pasquier, 2018] Pasquier, D. (2018). *L'internet des familles modestes. Enquête dans la France rurale*. Presses des Mines.

[Passmann *et al.*, 2014] Passmann, J., Boeschoten, T. et Schäfer, M. T. (2014). The gift of the gab : Retweet cartels and gift economies on twitter. *Twitter and society*, 331–344.

[Pearce *et al.*, 2014] Pearce, W., Holmberg, K., Hellsten, I. et Nerlich, B. (2014). Climate change on twitter : Topics, communities and conversations about the 2013 IPCC working group 1 report. *PloS one*, *9*(4), e94785.

[Powers, 2005] Powers, E. A. (2005). Interpreting logit regressions with interaction terms : an application to the management turnover literature. *Journal of Corporate Finance*, *11*(3), 504–522. http://dx.doi.org/10.1016/j.jcorpfin.2004.08.003. Récupéré le 2018-07-24 de http://www.sciencedirect.com/science/article/pii/S0929119904000628

[Prasetyo, 2014] Prasetyo, N. D. (2014). Tweet-based election prediction.

[Proulx, 2011] Proulx, S. (2011). *La puissance d'agir d'une culture de la contribution face à l'emprise d'un capitalisme informationnel : premières réflexions*. Culture et barbarie : communication et société contemporaine. Hommage à Edgar Morin. Athènes. 26-28 mai 2011.

[Queiroz *et al.*, 2018] Queiroz, G. D., Hvitfeldt, E., Keyes, O., Misra, K., Robinson, D. et Silge, J. (2018). *tidytext : Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools*. Récupéré le 2018-07-24 de https://CRAN.R-project.org/package=tidytext

[Quinn *et al.*, 2009] Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. et Radev, D. R. (2009). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228. http://dx.doi.org/10.1111/j.1540-5907.2009.00427.x. Récupéré le 2018-07-24 de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00427.x

[Raynauld et Greenberg, 2014] Raynauld, V. et Greenberg, J. (2014). Tweet, click, vote : Twitter and the 2010 ottawa municipal election. *Journal of Information Technology & Politics*, *11*(4), 412–434.

[Reynié, 2013] Reynié, D. (2013). *Les nouveaux populismes* (édition revue et augmentée éd.). Fayard/Pluriel.

[Robinson, 2018] Robinson, D. (2018). Who wrote the anti-trump new york times op-ed ? using tidytext to find document similarity. *Varianceexplained.org, http ://varianceexplained.org/r/op-ed-text-analysis/*.

[Roman et Bilan, 2012] Roman, A. et Bilan, I. (2012). The euro area sovereign debt crisis and the role of ECB's monetary policy. *Procedia Economics and Finance*, *3*, 763–768.

[Sabidussi, 1966] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

[Sanger, 2014] Sanger, W. (2014). *Valorisation de l'information sur les marchés financiers par l'utilisation des mégadonnées*. Polytechnique Montréal.

[Sanger et Warin, 2018a] Sanger, W. et Warin, T. (2018a). The 2015 canadian election on twitter : A tidy algorithmic analysis. International Conference on Computer Science and Computation Intelligence IEEE Xplore.

[Sanger et Warin, 2018b] Sanger, W. et Warin, T. (2018b). Jaccard similarity of 1517 european political manifestos across 27 countries (1945-2017). *Data in Brief, submitted.*

[Sanger et Warin, 2018c] Sanger, W. et Warin, T. (2018c). The public's perception of political parties during the 2014 québec election on twitter. *Canadian Journal of Communication*, *43*(2).

[Scott Long, 1997] Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, *7*.

[Silge et Robinson, 2016] Silge, J. et Robinson, D. (2016). tidytext : Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software*, *1*(3), 37.

[Silge et Robinson, 2017] Silge, J. et Robinson, D. (2017). *Text mining with R : A tidy approach*. " O'Reilly Media, Inc.".

[Small *et al.*, 2014] Small, T. A., Jansen, H., Bastien, F., Giasson, T. et Koop, R. (2014). Online political activity in canada : the hype and the facts. *Canadian parliamentary review*, *37*(4), 9–16.

[Smyth et Best, 2013] Smyth, T. N. et Best, M. L. (2013). Tweet to trust : social media and elections in west africa. Dans *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development : Full Papers-Volume 1*, 133–141. ACM.

[Snyder et Strömberg, 2008] Snyder, James M, J. et Strömberg, D. (2008). Press coverage and political accountability. `http://dx.doi.org/10.3386/w13878`. Récupéré le 2018-07-23 de `http://www.nber.org/papers/w13878`

[Song *et al.*, 2014] Song, M., Kim, M. C. et Jeong, Y. K. (2014). Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intelligent Systems*, *29*(2), 18–26. `http://dx.doi.org/10.1109/MIS.2014.20`

[Spinu *et al.*, 2018] Spinu, V., Grolemund, G., Wickham, H., Lyttle, I., Constigan, I., Law, J., Mitarotonda, D., Larmarange, J., Boiser, J. et Lee, C. H. (2018). *lubridate : Make Dealing with Dates a Little Easier*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=lubridate`

[Steiner, 1997] Steiner, J. (1997). *European Democracies* (4 edition éd.). Longman.

[Stockemer, 2018] Stockemer, D. (2018). The internet : An important tool to strengthening electoral integrity. *Government Information Quarterly*, *35*(1), 43–49.

[Strömberg, 2004] Strömberg, D. (2004). Radio's impact on public spending. *The Quarterly Journal of Economics*, *119*(1), 189–221. http://dx.doi.org/10.1162/003355304772839560. Récupéré le 2018-07-23 de https://academic.oup.com/qje/article/119/1/189/1876059

[Strömberg, 2015] Strömberg, D. (2015). Media and politics. *Annual Review of Economics*, *7*(1), 173–205. http://dx.doi.org/10.1146/annurev-economics-080213-041101. Récupéré le 2018-07-23 de https://doi.org/10.1146/annurev-economics-080213-041101

[Sullivan et Bélanger, 2016] Sullivan, K. et Bélanger, P. (2016). La cyberdémocratie québécoise : Twitter bashing,# VoteCampus et selfies. *Politique et Sociétés*, *35*(2), 239–258.

[Thabtah, 2008] Thabtah, F. (2008). VSMs with k-nearest neighbour to categorise arabic text data.

[Treiman, 2014] Treiman, D. J. (2014). *Quantitative data analysis : Doing social research to test ideas*. John Wiley & Sons.

[Tumasjan *et al.*, 2010] Tumasjan, A., Sprenger, T. O., Sandner, P. G. et Welpe, I. M. (2010). Predicting elections with twitter : What 140 characters reveal about political sentiment. Dans *Fourth International AAAI Conference on Weblogs and Social Media*. Récupéré le 2018-07-23 de https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441

[Vaccari *et al.*, 2013] Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J. et Tucker, J. (2013). Social media and political communication : a survey of twitter users during the 2013 italian general election. *Rivista italiana di scienza politica*, *43*(3), 381–410.

[van Eck *et al.*, 2006] van Eck, N. J., Berg, J. et van N. J. P, E. (2006). Visualizing the computational intelligence field. *IEEE Computational Intelligence Magazine - IEEE COMPUT INTELL MAG*, *1*, 6–10. http://dx.doi.org/10.1109/CI-M.2006.248043

[Warin *et al.*, 2014] Warin, T., de Marcellis-Warin, N., Troadec, A., Sanger, W. et Nembot, B. (2014). Un état des lieux sur les données massives. *CIRANO Rapport Bourgogne*.

[Warin *et al.*, 2018] Warin, T., Le Duc, R. et Sanger, W. (2018). Mapping innovations in artificial intelligence through patents : A social data science perspective. *Intelligence Conference of Computer Science and Computational Intelligence, IEEE Xplore*.

[Warin et Sanger, 2018] Warin, T. et Sanger, W. (2018). Connectivity and closeness among international financial institutions : A network theory perspective. *International Journal of Comparative Management*.

[Weller *et al.*, 2014] Weller, K., Bruns, A., Burgess, J. E., Mahrt, M. et Puschmann, C. (2014). Twitter and society : an introduction. In *Twitter and society*, volume 89 xxix–xxxviii. Peter Lang.

[Wickham, 2010] Wickham, H. (2010). ggplot2 : elegant graphics for data analysis. *J Stat Softw*, *35*(1), 65–88.

[Wickham, 2014] Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

[Wickham, 2016] Wickham, H. (2016). *ggplot2 : elegant graphics for data analysis*. Springer.

[Wickham, 2017] Wickham, H. (2017). *reshape2 : Flexibly Reshape Data : A Reboot of the Reshape Package*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=reshape2`

[Wickham *et al.*, 2018a] Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K. et RStudio. (2018a). *ggplot2 : Create Elegant Data Visualisations Using the Grammar of Graphics*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=ggplot2`

[Wickham *et al.*, 2018b] Wickham, H., François, R., Henry, L., Müller, K. et RStudio. (2018b). *dplyr : A Grammar of Data Manipulation*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=dplyr`

[Wickham *et al.*, 2018c] Wickham, H., Henry, L. et RStudio. (2018c). *tidyr : Easily Tidy Data with 'spread()' and 'gather()' Functions*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=tidyr`

[Wickham et RStudio, 2017a] Wickham, H. et RStudio. (2017a). *scales : Scale Functions for Visualization*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=scales`

[Wickham et RStudio, 2017b] Wickham, H. et RStudio. (2017b). *tidyverse : Easily Install and Load the 'Tidyverse'*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=tidyverse`

[Wickham et RStudio, 2018] Wickham, H. et RStudio. (2018). *stringr : Simple, Consistent Wrappers for Common String Operations*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=stringr`

[Xiao et Li, 2018] Xiao, N. et Li, M. (2018). *ggsci : Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'*. Récupéré le 2018-07-24 de `https://CRAN.R-project.org/package=ggsci`

[Yaqub *et al.*, 2017] Yaqub, U., Chun, S. A., Atluri, V. et Vaidya, J. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. *Government Information Quarterly*, *34*(4), 613–626.

[Zirn *et al.*, 2016] Zirn, C., Glavaš, G., Nanni, F., Eichorts, J. et Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. Dans *PolText 2016 : The International Conference on the Advances in Computational Analysis of Political Text : proceedings of the conference : sponsored by the European Social Fund, Operational Programme Efficient Human Resources 2014–2020*, 88–93. University of Zagreb. Récupéré le 2018-07-24 de `https://ub-madoc.bib.uni-mannheim.de/41552`

ANNEXE A ARTICLE 3

Table A.1 Marginal Effects of the binary logistic model

Dependent variable: topic {social OR integrity OR economy OR geopolitics}								
Model: binary logit	**Social**		**Integrity**		**Economy**		**Geopolitics**	
Independent variables	**Coef.**	**Marginal Effects**	**Coef.**	**Marginal Effects**	**Coef.**	**Marginal Effects**	**Coef.**	**Marginal Effects**
Jonathan	-0.0022344 ***	-0.0003645 ***	-0.0012095 **	-0.0002143 ***	0.0002237	0.0000392	0.0012038 ***	0.0002227 ***
Buhari	0.0003081	0.0000503	0.0009409 ***	0.0001667 ***	-0.0006147	-0.0001078	-0.0004339	-0.0000803
PDP	0.0029338 **	0.0004786 ***	-0.002108	-0.0003918 *	0.0000801	0.0000141	-0.0009846	-0.0001822
APC	-0.0044883 **	-0.0007322 **	0.0007484	0.0001326	-0.0036783 *	-0.000645 **	0.0031263 ***	0.0005784 ***
Constant	-0.7612005 ***		-1.157779 ***		-0.7943117 ***		-1.467405 ***	
Number of observations	540		540		540		540	
LR chi2	31.94		21.30		20.18		51.42	
Prob > chi2	0.0000		0.0003		0.0005		0.0000	
Pseudo R2	0.0526		0.0351		0.0332		0.0847	
Log likelihood	-287.68904		-293.00921		-293.57165		-277.95044	

P-value: *<0.1, **<0.05, ***<0.01

ANNEXE B ARTICLE 4

Table B.1 Pearson's product-moment correlation between Political Platforms

Country	Year	Political Party	Governing Party	PPMCC
Austria	2002	Freiheitliche Partei Österreichs	SPO	0.6579612***
Austria	2006	Freiheitliche Partei Österreichs	SPO	0.6672213***
Austria	2008	Freiheitliche Partei Österreichs	SPO	0.5854784***
Austria	2013	Freiheitliche Partei Österreichs	SPO	0.4384143***
Belgium	2003	Vlaams Belang	VLD	0.8747916***
Belgium	2007	Vlaams Belang	CDV	0.6678100***
Belgium	2010	Vlaams Belang	NVA	0.5885338***
Bulgaria	2009	ATAKA	GERB	0.6921402***
Bulgaria	2013	ATAKA	GERB	0.9268929***
Bulgaria	2014	Patriotic Front	GERB	0.6808393***
Bulgaria	2014	ATAKA	GERB	0.6719154***
Bulgaria	2017	United Patriots	GERB	0.6456762***
Croatia	2000	Hrvatska stranka prava	SDP	0.4264476***
Croatia	2003	Hrvatska stranka prava	HDZ	0.5544504***
Denmark	2001	Dansk Folkeparti	V	0.2540310***
Denmark	2005	Dansk Folkeparti	V	0.0066146
Denmark	2007	Dansk Folkeparti	V	0.1575774
Denmark	2011	Dansk Folkeparti	V	0.7740050***
Estonia	2015	Eesti Konservatiivne Rahvaerakond	ER	0.8030181***
Finland	2003	Perussuomalaiset	SK	0.4737018***
Finland	2007	Perussuomalaiset	SK	0.4269633***
Finland	2011	Perussuomalaiset	KK	0.7696836***
France	2002	Front National	UMP	0.7860805***
France	2012	Front National	PS	0.8482165***
France	2017	Front National	LREM	0.7320784***
Germany	2013	Alternative für Deutschland	CDUCSU	0.3464566***
Germany	2017	Alternative für Deutschland	CDUCSU	0.7081878***
Greece	2012	Chrysi Avgi	ND	0.9297183***
Greece	2015	Chrysi Avgi	SYRIZAEKN	0.8403733***
Hungary	2014	Jobbik Magyarországért Mozgalom	FiDeSz	0.7510525***
Italy	2013	Fratelli d'Italia	M5S	0.5842131***
Italy	2013	Lega Norte	M5S	0.6249190***
Latvia	2006	Tevzemei un Brivibai/LNNK	TP	0.3184229*
Latvia	2010	Nacionala apvieniba	U	0.5625352***
Latvia	2011	Nacionala apvieniba	SC	0.7896515***
Latvia	2014	Nacionala apvieniba	SDPS	0.5565748***
Netherlands	2006	Partij voor de Vrijheid	CDA	0.6557068***
Netherlands	2010	Partij voor de Vrijheid	VVD	0.6450216***
Netherlands	2012	Partij voor de Vrijheid	VVD	0.5083280***
Slovakia	2006	Slovenská národná strana	Smer	0.4991039***
Slovakia	2010	Slovenská národná strana	Smer	0.4847133***
Slovakia	2012	Slovenská národná strana	Smer	0.4475110***
Slovenia	2004	Slovenska Nacionalna Stranka	SDS	0.5427495***
Slovenia	2008	Slovenska Nacionalna Stranka	SDS	0.3520233***
Slovenia	2011	Slovenska Nacionalna Stranka	SDS	0.7019773***
Sweden	2010	Sverigedemokraterna	SAP	0.4968909***
Sweden	2014	Sverigedemokraterna	SAP	0.5457784***
UK	2001	UK Independence Party	Labour	0.6075615***
UK	2015	UK Independence Party	Conservatives	0.8669829***

Note: Significance level: *** 99%, ** 95%, * 90%, PPMCC: Pearson's product-moment correlation coefficient